The Effects of Different Types of Textual Input Enhancements on Incidental and Intentional Vocabulary Learning From Reading

Sauer, Bianca Maria

Awarding institution:
King's College London

The Effects of Different Types of Textual Input Enhancements on Incidental and Intentional Vocabulary Learning From Reading

Bianca Sauer

A Thesis Submitted In Fulfilment Of The Requirements For A Phd Degree At King's College London

School of Social Science and Public Policy

Department of Education and Professional Studies

King's College, London

September 2016

ABSTRACT

This study investigates the effects of three kinds of textual input enhancements (TIEs) - bold-printing, L2 glossing, and a combination of the two - on tasks aimed at facilitating incidental and intentional vocabulary learning from reading. It explores which other task-related and learner-internal factors influence vocabulary acquisition. Previous research on vocabulary learning from reading found positive effects for the provision of enhancements. However, findings are inconclusive regarding which types of enhancements are most effective for which type of vocabulary knowledge, and there has been little research investigating the effects of 'obtrusiveness', i.e. the interruption of the reading flow through consulting glosses. Likewise, few studies have considered how learners interact with TIEs.

269 Danish secondary school L2 learners of English participated in three reading/testing sessions. They read either unmodified texts (control group) or texts in which target words were highlighted in the three enhancement forms. Immediate active and passive recognition and passive recall of target word meaning were assessed in a vocabulary post-test. Volunteers participated in retrospective interviews.

To compare the effects of incidental and intentional word learning, for the analysis the data were split into those collected after the first session, where no focus on vocabulary learning was assumed, and subsequent reading/testing sessions, where learners increasingly focussed on vocabulary learning. Correlation computations confirmed the assumed relationship between TIE-use and vocabulary acquisition. The results concerning the impact of the different TIE types varied from session to session, but showed that enhancement use of any type had the greatest impact on establishing a form-meaning link measured in a receptive meaning recall test. Regression calculations revealed that variables such as testing session or text type significantly predicted the outcomes of the vocabulary post-test.

Even though many interviewees perceived 'obtrusiveness' as problematic, the enhancement types involving glosses led to significantly higher vocabulary post-test scores than bold-printing only. The interviews suggest that especially the enhancement type which combined bold-printing and glossing encouraged learners to focus on the target words in ways that initiate deep processing. Bold-printing of target words, however, often procured results that were similar to those from reading unenhanced texts. Several interviewees found working with such typographic enhancements 'confusing'.

The interviewees described behaviour specific to the different TIE types. Their general TIE approach seemed habit-driven, economical, and related to the cognitive involvement load factor '*need'*. The statistical analysis and the interviews showed that repeated testing had an effect on how learners approached the tasks. These findings shed light on the complexity of the relationship between incidental and intentional word learning and on how research procedures can influence outcomes.

The outcomes confirm the usefulness of enhancements for vocabulary learning. However, they also show that great care has to be taken when providing TIEs for any language learning purpose, as learner behaviour related to their application is far more complex, and therefore deserves more consideration, than is currently given.

Acknowledgements

I dedicate this thesis to
Johannes and Bjørk
△
For your love and patience

First and foremost, I would like to extend my gratitude to my competent, understanding, hard-working, and always encouraging supervisors, Ursula Wingate and Nick Andon. Thank you for your unremitting patience! I could not have asked for better mentors on this long and, at times, arduous journey.

A special thanks to the many students and teachers in Denmark for dedicating their time and sharing their thoughts and knowledge with me; at Stenhus Gymnasium, Falkonergårdens Gymnasium, Ørestad Gymnasium, and Greve Gymnasium. I am particularly grateful to Christoffer, Hanne, Marie, Marianne, Merete, Nikolaj, Owen, Søren, Susanne, Tania, and Trine.

I also benefited greatly from the support of and enlightening discussions with my fellow PhD students, in particular Sophie Kelsall, Taehee Choi, and Emily Dawson.

I am deeply indebted to Birgit Henriksen, who so graciously shared her time and extensive expertise with me. I am grateful to Parvaneh Tavakoli, for being a beacon of hope in difficult times. Thanks are also due to Joyce Kling Sørensen, Kimberley Chopin, Jens Bahns and Henning Wode.

I would also like to thank Celia Roberts, Ben Rampton, and Constant Leung for their support with the GradEast Graduate School in Linguistics scholarship at Copenhagen University.

This thesis is written in memory of my supervisor Johannes Eckerth, who was an inexhaustible source of inspiration, wisdom, and encouragement during the initial stages of this project.

INDEX

LIST OF TABLES

LIST OF FIGURES

LIST OF TABLES, FIGURES, AND EQUATIONS IN THE APPENDIX

TABLES IN THE APPENDIX

EQUATIONS IN THE APPENDIX

FIGURES IN THE APPENDIX

1. INTRODUCTION

2. LITERATURE REVIEW

3. METHODOLOGY CHAPTER

1. INTRODUCTION

1.1. Background

As a foreign language teacher, I find that learners are often expected to know how to read effectively and to have all the necessary resources and strategies at their disposal. One of these resources is a sufficiently large vocabulary. Research has shown that second-language readers rely heavily on vocabulary, that lexis-related problems are a dominant source of difficulty (Révész & Brunfaut, 2013), and that a lack of vocabulary knowledge in particular is "the largest obstacle for second-language readers to overcome" (Huckin & Bloch, 1993:154; see also Alqahtani, 2015; Vermeer, 2001). However, due to lack of time, resources and sometimes a lack of awareness of the importance of vocabulary, many foreign language courses provide insufficient support for the learning[1] of new words. Cobb (1999:345) pointed out that "students typically need to know words measured in thousands, not hundreds, but receive language instruction measured in months, not years". Some theoretical approaches to L2 vocabulary acquisition propose that instead of being exposed to explicit L2 vocabulary teaching, students should read extensively, as comprehension-focused reading is assumed to induce vocabulary acquisition (Nation, 2001). This is in line with acquisition oriented approaches such as TBLT, CLIL and strong versions of CLT, which are underpinned by the principle that all aspects of language can be acquired implicitly and subconsciously during meaning-focused activities, particularly if learners' attention can be drawn to form at some stage in the process. Proponents of these approaches argue that the majority of vocabulary is learned receptively through reading (or listening, N. Ellis, 1995; R. Ellis, 2003; Nagy et al., 1985; Schmitt, 2010; Webb, 2005). Learners who have reached a certain level of language proficiency, often rely on reading as a source of those words (Schmitt, 2010).

Reading is highly valuable for encountering and learning new words in authentic contexts and beneficial for language learning overall (Grabe, 2009a). However, acquiring new vocabulary through reading is comparatively slow and ineffective (Horst, et al., 1998; Laufer, 2003; 2005; Min, 2008; Nation, 2001). As proficient L2 readers are able to draw on their schematic knowledge to understand a text without understanding every word, they typically ignore a vast number of words or other language forms they encounter while reading and are generally insensitive to specific features in target language input. They might skip words they do not find relevant. Also, learners might not notice the correct word- or grammatical forms due to poor input characteristics of specific items, or because of a lack of noticing ability (Sharwood Smith, 1993). Therefore, research into ways to draw learners' attention to lexical

[1] In this study, the terms 'learning' and 'acquisition' are used interchangeably.

items and other language features, with the potential to increase noticing and eventually acquisition of target words is of great interest. Furthermore, research that examines which word interventions increase the rate of word learning through reading and investigates how this objective can be achieved more effectively is also important (Cho, 2010; Peters et al., 2009; Schmitt, 2008; Zahar, Cobb, & Spada, 2001).

1.2. Setting The Scene: Key Terms And Questions

When I started my PhD project, I was intrigued by new research that had just been published about the 'involvement load hypothesis' (Keating, 2008; Kim, 2008). As a teacher I saw great potential in the idea that vocabulary learning could be made more effective by designing tasks that carry a high level of 'cognitive involvement load', i.e. lead to deep processing of vocabulary items (Hulstijn & Laufer, 2001; Laufer & Hulstijn, 2001). What surprised me, however, was that the studies lending support to this hypothesis used 'textual input enhancements' (TIEs) without considering the use of the enhancements as a learning process that demanded 'cognitive involvement'. I therefore decided not to focus my research on the involvement load hypothesis, but instead to investigate how TIEs are used and how they influence vocabulary learning.

Sharwood Smith (1991:118) defined TIEs as "consciousness-raising" through artificially and externally induced salience. They are a complex, extensively researched phenomenon (e.g. Boers et al., 2016; Han, Park, & Combs, 2008; Ko, 2012; LaBrozzi, 2016; O'Donnell, 2012; Rott, 2007). In SLA research, enhancements are regarded as a relatively implicit and less burdensome technique than others for drawing learner attention to both form and meaning (Lee, 2007; Roby, 1999). It is assumed that enhancing texts generally helps learners to understand authentic texts better and to learn the language along the way. This is because input must be noticed and understood to be processed for intake. Enhancing texts for the purpose of facilitating noticing is based on the necessity to distribute learners' attention. There is an ongoing scholarly debate on the amount of attention that learners have available (Pienemann, 2003; Robinson, 2001; Skehan & Foster, 2001). However, there is largely agreement on the fact that attention and input processing capacity are somehow limited. Therefore, guiding learners' attention to desirable language forms may be useful to boost learning. Enhancements can raise salience and make noticing more likely. The more salient the target words and the more the learners are cognitively involved in making sense of them, the more likely is initial and mid-term acquisition and retention of these words (Baddeley, 1997; Han et al., 2008; Laufer & Hulstijn, 2001; Sharwood Smith, 1993). Therefore it is assumed that

"lexical enhancements and intervention tasks, which guide readers' attention to specific words, increase the likelihood for word gain" (Rott, 2007:167). In language learning, salience refers to the ease with which a linguistic term is perceived (Schmidt, 1990; Svensson, 2012). Research has found that aspects like for instance frequency, relevance within context, learner motivation, and word form play a role in vocabulary acquisition.

Various types of input enhancements have been researched within second language acquisition, including glosses, gestures, repetition and slowing down the rate of speech, input flood in reading, or visual highlighting of words, but the results give an inconclusive picture. This is due to the great variety of investigated enhancement types, used methodology and involved variables (see LaBrozzi, 2016 for an overview). In my study, I focus on TIEs, i.e. manipulations found in written input for reading. TIEs are defined in this study as any intentional changes to specific items in the textual input learners receive with the purpose of aiding learning and/or comprehension by raising the salience of these items. In my study, I investigate the use of TIEs used to serve both purposes, to support vocabulary acquisition and to aid comprehension of texts. The primary focus is, however, on vocabulary acquisition.

TIEs are often lexical enhancements, such as L1 or L2 glosses in the margin or in attached vocabulary lists, but may also take the form of typographic changes, such as bold-printing or underlining target items. Another form of TIE is input flood, where salience is created by (usually, artificially engineered) frequency of exposure to language items (Pacheco, 2004; White, 1998). Language learners are directed to read extensively and are thus exposed to vast amounts of these language items, for instance unknown vocabulary. Processing a new word repeatedly in one or multiple texts has been found to be conducive to incidental word learning (Horst et al., 1998; Rott, 2007). Lately, the increased use of modern technology devices in language classrooms has led to new ways of enhancing reading materials, for instance through internal dictionaries, pictures, videos, or activities linked to target items (Lomicka, 1998; Webb, 2012; Abraham, 2008).

Although one of the main functions of enhancements is to raise salience, the different factors which affect whether words really appear as salient and are processed further have not been extensively researched (Izumi, 2002; O'Donnell, 2012). For example, it is not clear how particular types of enhancements link to specific aspects of language acquisition. This issue is important because different types of enhancements involve cognitive procedures that differ in terms of complexity and time expenditure (Gettys, 2001). Glosses, for instance, may not be helpful for promoting long-term word retention as they deprive learners of the opportunity for deeper processing through inferring.

Other factors may affect the noticing and acquisition of vocabulary through reading enhanced texts. These may be related to the task or to students' learning preferences. Investigating the latter is especially important to examine, as externally created salience (e.g. by a teacher) does not necessarily result in internal salience (Sharwood Smith, 1993). The way learners work with texts determines whether enhancements are effective.

Another question to investigate is whether vocabulary knowledge derived from TIEs is retained in the longer term. Furthermore, given that enhancements are intended to compensate for lack of time in vocabulary teaching, investigations into which facets of vocabulary knowledge (e.g. receptive/passive or productive/active knowledge) can be acquired by reading enhanced texts are important. Thus, research is needed into whether enhancements support passive or active knowledge and for how long this knowledge is retained, as long-term retention of learned vocabulary is the ultimate goal.

1.3. Research Objectives

In order to provide a comprehensive view of the above-mentioned problems, this research project is both hypothesis testing and contributing to the data base. It aims to expand the current understanding of initial word learning and long-term word acquisition by exploring how it is affected by four overlapping variables: types of textual input enhancement, learner behaviour, task design, and learning approach.

I chose to use three different types of enhancements in my reading tasks; bold printing, L2 glossing, and a combination of the two. These were chosen for three reasons: they are the most commonly used types of enhancements in Danish EFL-teaching materials, they differ in their focus on either word form, meaning, or both, and, as they might encourage readers more or less strongly to reflect unknown vocabulary, they seemingly influence the reading flow to different degrees (Bell & LeBlanc, 2000; Rott, 2007, Simard, 2009). It therefore seems important to find out whether there is a hierarchy to enhancement types, which is reflected in their effect on vocabulary acquisition, i.e. whether less intrusive forms of enhancement (e.g. bold printing) aid acquisition to a lesser degree than seemingly more intrusive forms of enhancement (e.g. glossing). Another important question was whether different types of enhancements foster different facets of vocabulary knowledge skills, for instance active and passive recall and recognition. Moreover, I believed that the efficacy of TIE for vocabulary learning depends on the type of reading task and on how language students use the enhancements in them. I therefore wanted to find out what learners actually do with the enhancements.

Accordingly, the objectives of this study were to investigate the following issues:

- Do textual input enhancements in the form of L2 glosses and/or bold printed target words have a beneficial effect on vocabulary learning from reading?
- How do learners adapt reading tasks containing enhancements to their individual learning preferences?
- In what way does learner behaviour have an effect on vocabulary learning?
- How does a learner's approach to textual input enhancements relate to incidental or intentional vocabulary learning?
- Do different types of enhancements have a specific impact on various types of vocabulary learning?

These issues are brought together in the three research questions to be answered in this study:

(1) What are the immediate and long-term effects of different types of textual input enhancement on incidental and intentional learning of EFL-vocabulary from reading tasks?
(2) Does textual input enhancement have a differential effect on different types of word knowledge?
(3) Which task-related[2] and learner-internal[3] factors emerge as relevant for vocabulary learning from reading enhanced texts and how do these affect how learners use input enhancements in reading tasks?

In order to answer these questions, this study goes beyond previous research in four ways: Firstly, it complements quantitative data with qualitative data (retrospective semi-structured interviews), and therefore provides evidence on possible influences of TIE-related learner behaviour on word learning. Secondly, it is based on data gathered from secondary school EFL-learners, whereas most previous studies researched university learners. Hence, its results may be useful for the wide population of school teachers and learners. Thirdly, this study focuses on receptive and productive aspects of vocabulary knowledge. Finally, I applied a non-experimental research design (based in authentic classrooms), whereas the bulk of related research is carried out under experimental conditions.

[2] Factors directly relating to the task, e.g. task instructions, task structure.
[3] Factors directly relating to learners, e.g. learning habits.

1.4. Outline Of My Study

Drawing on previous research and theories into second language vocabulary acquisition, chapter two (Literature Review) first discusses what it means to know a word and how second language vocabulary is acquired. Next, implicit vocabulary acquisition is discussed with a particular focus on acquiring vocabulary through reading and the role of teaching. Following this, research related to TIEs is examined in detail. The chapter concludes by identifying areas that require further research relevant to my own study. Chapter three starts with a discussion of the general research approaches selected to investigate my research questions. It provides a detailed description and justification of the specific methods and instruments used in this research, as well as the approach to data analysis.

As two types of data were gathered, the outcomes of their analysis are presented in separate chapters. Chapter four (Quantitative Findings) starts with a presentation of the hypotheses underlying my research approach. Then, the findings of the three reading/testing sessions are reported successively. The test results are, however, also analysed as a whole. The outcomes of this analysis are presented then. Finally, the results of the delayed vocabulary post-test are revealed. The Qualitative Findings chapter is structured according to different themes which emerged in the interview data analysis. Four overarching themes 'textual input enhancements', 'recalling vocabulary knowledge in the interviews', 'reading and vocabulary strategies', and 'task, text and test-related issues', head the subsections, and disclose how TIEs affected vocabulary learning and which other factors had an impact on the word acquisition processes.

In the Discussion, the findings of both types of data are brought together and discussed in relation to previous research in the field. New insights gained from the analysis of my data are presented. The Conclusions sum up the major outcomes, point to the significance of these findings for teachers, discuss the limitations of my study, and make recommendations for future research.

The details of the research design used in this study are illustrated in Table 1 below:

Table 1: Research Design Procedure Illustration

	Objectives	Procedure
Pre-Test I (N=45)	- Finding suitable texts and target words - Finding suitable pre- and post-reading exercises - Designing the most fitting test	- Potential target words and distractor words presented decontextualized in lists - Participants were asked to report their (degree of) knowledge of these words in table adapted from the vocabulary knowledge scale test[4]

[4] See Table 4

Pre-Test II (N=62)	- Narrowing down the chosen texts and target words - Checking the suitability of the test design	- Participants read the chosen texts, which contained the unenhanced target words - Participants were asked to highlight all unfamiliar words - Participants worked through list with potential target words and distractors: translated or used words in full sentence - Participants indicated how 'interesting and relevant' they had found the chosen texts[5]
Pilot Study (N=55)	- Ensuring the feasibility and validity of the research design - Testing the interview guide questions - Checking suitable time on task for reading / testing sessions	- Participants completed reading task in various conditions, including pre- and post-reading exercises - The vocabulary post-test was completed - Test interviews conducted with two volunteers
		Week I: Session 1 - Reading across texts and TIE types; e.g. class 1: text 1 – reading condition 1 (bold-printed target words); class 2: text 2 - reading condition 1 … - Retrospective interview 1 with 1-2 volunteers from each class
		Week II: Session 2 - Reading across texts and TIE types; e.g. class 1: text 2 – reading condition 2 (glossed target words); class 2: text 1 – reading condition 3 … - Retrospective interview 2 with the same 1-2 volunteers from each class
		Week III: Session 3 - Reading across texts and TIE types; e.g. class 1: text 3 – reading condition 3 (glossed and bold-printed target words); class 2: text 3, reading condition 2 … - Retrospective interview 3 with the same 1-2 volunteers from each class
Delayed Post-Test (N=107)	- Conducted 4 weeks later - Gauging vocabulary knowledge attrition	- Participants completed a vocabulary post-test comprising of all 30 target words presented in list

[5] See Figure 2
[6] See Table 3 for further details

2. LITERATURE REVIEW

2.1. Chapter Outline

I start this chapter with an outline of research concerning basic concepts of vocabulary knowledge (2.2). I explain what 'knowing a word' might mean, and which different forms of word knowledge there are. Furthermore, I discuss the literature on acquisition of vocabulary knowledge (2.3). The focus is on how attention distribution and awareness are linked to vocabulary acquisition, and which vocabulary acquisition theories contribute to understanding the effects of textual input enhancements (TIEs). In 2.4, I discuss research regarding the learning and teaching of vocabulary in foreign language classrooms, with a specific focus on vocabulary learning strategies. I then discuss the specific features of L2 reading processes and research on how vocabulary is learned through reading (2.5).

Finally, in section 2.6, I review previous TIE studies. The section starts with an outline of research into the effects of different types of TIE and other foci in the field. Building on this, I conclude this section by addressing open questions within the research corpus and with research questions that result from the literature review and describe which contribution my study tries to make to the field.

2.2. Understanding Vocabulary Knowledge

To get a better understanding of how input enhancements might help language learners to acquire new vocabulary through reading, it is important to consider what vocabulary knowledge is. The following sections are dedicated to this purpose.

2.2.1. What It Means To Know A Word

The term 'word' is rather elusive. The question of what exactly might be regarded as a 'word' has been discussed intensely. Various fields of linguistics offer diverse perspectives on the matter and lead to different definitions. Even though some basic criteria have been agreed on, to describe the very heterogeneous concept of 'word' linguistically (acoustic and semantic identity, morphological stability, syntactic mobility), none of these offer a definition which would cover all the aspects of what a word is. 'Word' remains a not clearly definable linguistic unit.

In this study, no new attempt was undertaken to find a universally valid definition. Instead, the concept 'word' was based on the context of language learning within a classroom. According to Bloom (2000:16), "the relevant sense of *word* from the standpoint of language

acquisition should include all and only those forms whose meanings must be learned". Therefore, as in previous studies (Eckerth & Tavakoli, 2012; Laufer et al., 2004; Pellicer-Sánchez & Schmitt, 2010), productive and receptive word knowledge will be distinguished and only the tested facets of word knowledge will be considered.

The term 'word', is often used interchangeably with 'vocabulary' to describe the contents of language. Just like with 'word', the distinction between productive (active) and receptive (passive) vocabulary knowledge seems to be more important than a distinct understanding of what vocabulary is (Crystal, 2003). In his widely quoted book on vocabulary learning research, Nation (2001) does not specifically explain 'word' and 'vocabulary', but rather points out the differences between tokens (running words) and types (different kinds of words) and lemmas (headwords) within vocabularies. He highlights that words are organized in *lexical word families*, while the term lexicon highlights structure and scope of vocabulary. Nation principally understands vocabulary as the plural form of *word*, i.e. the words of a language are its vocabulary. In a different approach, Gass (1999) refers to the common understanding that "learning a second language means learning its vocabulary" (p. 325). Neither of the two concepts 'word' and 'vocabulary' is sufficient in today's understanding. It is often more precise to refer to 'lexical items' rather than words, as most units of meaning consist of more than one word, for instance collocations or lexical bundles (Bahns, 1993; Schmitt, 2000; Tseng & Schmitt, 2008).

For the purposes of my study, vocabulary is best defined regarding the concept of word knowledge.

Trying to define word knowledge is, however, as difficult as defining a 'word', because "there are many degrees of knowing" (Nation, 2001:23), and many facets to a word: a word has multiple shades of meaning and more than one form. A learner needs to know how to use a word, how it performs in a sentence, and its relationship to other words. There is also the issue that it is not enough to know individual words but that lexical items can include multi-word units, idioms, and phrasal verbs. Complete knowledge of a word would have to cover all this syntactical, semantic, pragmatic, phonological, orthographic, morphologic information and is therefore rarely achieved, even by native speakers. Learners, researchers, and teachers have to be aware of the state of "fuzziness" of word knowledge (Thornbury, 2002:16). Modelling word knowledge is therefore a challenging task and there are different views as to how detailed the modelling should be in order to be practical. Meara proposed a three-dimensional model in which he applied size, organization, and accessibility as categories (Meara, 1996). Henriksen (1999) differentiated between the three categories of 'partial-precise-' (ranging from mere recognition to precise comprehension), 'receptive-productive' (ranging from receptive to

productive mastery of a word), and 'depth knowledge' (ranging from a lexically loosely- to well-integrated word). Richards (1976) identified several features of word knowledge; frequency, register, form, associations, meaning-concept, and meaning-associations. Nation extended Richards' list to nine types of word knowledge; each specified both for receptive and productive knowledge (Nation, 1990; 2001). As Nation's model of word knowledge is one of the most encompassing attempts to illustrate the different dimensions of word knowledge, it is reproduced below:

Table 2. Nation's Word Knowledge Model (Nation, 2001:27)

What is involved in knowing a word			
Form	spoken	R	What does the word sound like?
		P	How is the word pronounced?
	written	R	What does the word look like?
		P	How is the word written and spelled?
	word parts	R	What parts are recognizable in this word?
		P	What word parts are needed to express the meaning?
Meaning	form and meaning	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	concept and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	associations	R	What other words does this make us think of?
		P	What other words could we use instead of this one?
Use	grammatical functions	R	In what patterns does the word occur?
		P	In what patterns must we use this word?
	collocations	R	What word or types of words could occur with this one?
		P	What words or types of words must we use with this one?
	Constraints on use (register, frequency, …)	R	Where, when, and how often would we expect to meet this word?
		P	Where, when and how often can we use this word?
Note: R = receptive knowledge, P = productive knowledge			

The model presents form, meaning, and use as the overarching facets of word knowledge, each of which contains three subcategories (spoken form, written form). Different from Richards (1976), Nation distinguishes between receptive and productive knowledge for each of the features. Thus, the model illustrates the multi-factorial, incremental nature of vocabulary acquisition. It makes sense to conceive of lexical knowledge as a progressive scale rather than an either-or phenomenon, especially because vocabulary knowledge is subject to constant change (Klapper, 2006). Learning a word is most definitely not a one-off event and Schmitt (1998) even claims that there is not one single moment when we can say we fully 'know' a word.

Therefore, when the term 'acquisition' is used in my study it is never meant to denote full knowledge. However, the term is still useful, as it covers a wide range of vocabulary knowledge facets in SLA literature. However, the reader needs to be aware that in the context

of my study, acquisition likely describes merely initial stages of knowledge formation. Encountering new vocabulary once while reading is most likely to activate initial knowledge stages[7] only and is not likely to lead to deep processing, even when enhanced (Jiang, 2000).

While the definitions of word knowledge proposed above differ widely, there seems to be some agreement now that the receptive/productive and depth/breadth distinctions of word knowledge may be useful. What exactly these terms refer to and how they can be measured, however, is up for debate. This debate is described in the following sections.

2.2.2. Depth And Breadth Of Vocabulary Knowledge

Vermeer (2001) suggested that depth and size of vocabulary knowledge essentially may be the same construct. However, considering the existing research, it does not become quite clear whether they are distinct or even what exactly these terms refer to. It seems that how they are understood and researched very much depends on how they are conceptualized and measured. Schmitt's 2014 review of research on the relationship between vocabulary size and depth showed that those conceptualisations vary greatly.

The category of breadth of vocabulary knowledge, or vocabulary size, is often used to refer to "the number of words for which the person knows at least some of the significant aspects of meaning" (Anderson & Freebody, 1981:92f). It is a category that has been researched thoroughly (Aitchison, 1994; Clark, 1993; R. Ellis, Tanaka, & Yamazaki, 1994; Gass, 1988; Meara, 1996; Nation, 1990, 1993, 2001; Read, 2000). In testing, breadth is usually determined by the degree to which learners are able to link form to meaning (Webb, 2012). At a closer look, however, it is a somewhat vague category, as it does not distinguish between degrees of knowledge. The way in which vocabulary breadth has been determined means that a partially known word is registered as 'known' just as much as a word for which more comprehensive knowledge exists (e.g. with rich lexical and grammatical links, links to pragmatic knowledge). The problem seems to be that the quality of word knowledge is disregarded. Therefore, researchers started focussing on 'depth' of vocabulary knowledge, which considers just this quality (Haastrup & Henriksen, 2000; Henriksen, 1999; Meara, 1996; Nassaji, 2004; Read, 1993, 2000; Vermeer, 2001). According to Meara (1996), it became increasingly accepted that alongside measures of vocabulary size, independent measures to describe quality of word knowledge were needed. Gass and Selinker (2008) suggested that

[7] Initial knowlege, i.e. partial, probably receptive knowledge (Haastrup & Henriksen, 1998; Schmitt, 2010)

breadth of vocabulary knowledge was only relevant when accompanied by depth of knowledge.

As, according to Schmitt (2014) “virtually all aspects of vocabulary knowledge seem interrelated”, is is difficult to clearly define vocabulary knowledge depth and it has been conceptualized in many different ways (p. 92). There seems to be no definition of vocabulary depth that is widely agreed upon. However, the most common conceptualization of vocabulary depth has been lexical organization (e.g. Laufer & Goldstein, 2004; Tseng & Schmitt, 2008), with the two Word Associate Format (WAF) tests developed by Read (1993 and 1998 versions) used in many of these studies. Nation’s multidimensional (2001) model, described above, is one account of vocabulary depth: Depth of knowledge on the one hand describes how many aspects in Nation’s list are known, but also how well each aspect is known. Vermeer (2001) stated that it was “all about the fact that specific words can be known to a greater of lesser extent” and that “the denser the network around a word, the richer the set of connections around the word, the greater the number of words known, and the deeper the knowledge of that word” (p. 218). Depth of vocabulary knowledge comprises different forms of multidimensional knowledge, such as linguistic knowledge, world knowledge and strategic knowledge, i.e. knowing a word and its antonyms, spelling, collocations, associations, means that this word will be known more deeply, than a word that is only loosely integrated into a semantic field (Nassaji, 2004).

According to Henriksen (1999), in addition to ‘partial / precise’, and ‘receptive / productive’, depth forms the third of three levels of lexical knowledge. She found that depth shows that the way in which learned words are integrated into a dense lexical network illustrates the qualitative difference of how well vocabulary can be known. This qualitative dimension of language competence is linked to the process of network building that takes place in vocabulary learning (Haastrup & Henriksen, 2000; see section 2.3). Still, there is no one comprehensive measure of vocabulary knowledge depth. This is “partly due to depth being a very broad construct that cannot simply be measured in a single test or even practically in a battery of tests” (Schmitt, 2014:921).

Schmitt concludes from his 2014 review of research into vocabulary knowledge size and depth that these concepts are, albeit being independent knowledge components, related to a degree that they cannot be clearly distinguished. However, according to Schmitt, it remains unclear how exactly size and depth relate. His review showed that the understanding of the nature of the two concepts decisively depends on how they are investigated. Previous studies have, for example, measured size and depth of vocabulary knowledge in terms of the mastery of the form-meaning link at recognition and recall levels of mastery. Other studies investigated them

as knowing single- or multiple word knowledge aspects, examined the knowledge of collocations, or studied the ability to use lexical items fluently. This great variety of approaches makes definite conclusions difficult.

However, the distinction between size and depth can be relevant in practical terms. Schmitt (2014) suggests that awareness of size and depth can be valuable for "practitioners to drive home the need for rich, sustained instruction and input in order to develop knowledge beyond the simple memorization of the form–meaning links" (p. 942). It can further be useful for teachers to consider vocabulary depth in their planning. Acquiring sufficient word knowledge may often be a long process. Therefore, Webb suggests that "awareness of how different types of learning may affect vocabulary knowledge can help teachers to design and sequence activities to promote depth" (2013:1). Nation's (2001, 2007) four strands of teaching/learning activities (meaning-focused input, meaning-focused output, language-focused learning, fluency development, see section 2.4.1) can also provide a useful approach to developing vocabulary depth.

Depth of vocabulary knowledge is often conceptualized with respect to what learners are able to do with a lexical item, i.e. whether they have achieved receptive or productive mastery. Also Nation's word knowledge model, presented earlier, shows that each of the nine knowledge categories can be known receptively and productively. This is highly relevant for understanding the effects of TIE, as it is possible that only specific facets of word knowledge are activated through reading enhanced materials. Research into the underlying notions of the receptive/productive distinction is reviewed in the following section.

2.2.3. Receptive And Productive Vocabulary Knowledge

In addition and connected to the depth/breadth distinction, when conceptualizing lexical competence, the 'active/passive' or "'receptive/productive' scale of knowledge is one of the […] major ideas explored" (Nation, 2001:23). However, there still is a "lack of an accepted conceptualization of what receptive and productive mastery of vocabulary entails" (Schmitt, 2014:923). Commonly receptive/passive knowledge is associated with listening and reading and comprehension and perception, and productive/active knowledge on the other hand is linked to speaking and writing (e.g. Laufer, 2005; Webb, 2012). This distinction is closely related to understanding the scope of what it means to know a word: receptive knowledge is demonstrated for instance through being able to recognize a word when it is heard or knowing what the word means in the particular context. Productive knowledge, on the other hand, presents itself through recall activities like being able to write it with correct spelling or being able to say the word with correct pronunciation.

Word learning studies have shown that scores on productive knowledge are much lower than on receptive tests (Eckerth & Tavakoli, 2012; Laufer & Goldstein, 2004; Laufer & Paribakht, 1998; Pellicer-Sánchez & Schmitt, 2010; Webb, 2008) and that productive knowledge develops slower (Laufer, 2005; Webb, 2008). There is an assumption that receptive knowledge precedes productive knowledge. However, the latter may only be true if individual aspects of vocabulary knowledge rather than the broader construct of knowing words are considered (Webb, 2012). A learner may well be able to spell a word correctly (productive knowledge of word form) before having understood its correct meaning (receptive knowledge of form and meaning). Therefore, it is important to "differentiate between receptive and productive knowledge of individual aspects of knowledge rather than receptive and productive knowledge of words as a whole" when investigating these two dimensions (Webb, 2012).

Even though most language learners will confirm "that receptive learning and use is easier than productive learning and use, […] it is not clear why receptive use should be less difficult than productive use." (Nation, 2001:28). Several arguments might explain this issue: One is that, active knowledge is more difficult to achieve, because there is more to learn about a word to use it correctly (pronunciation, spelling, collocations, register), and another one is that most learners practice words less than they encounter them in the input (Laufer, 2005).

Nation claims that "when applied to vocabulary, these terms (receptive and productive) cover all the aspects of what is involved in knowing a word." (Nation, 2001:26). Taking this approach and the overall complexity of vocabulary knowledge into account it is not quite clear whether it makes sense to regard receptive/productive as a distinction at all. It has been suggested to better see receptive/productive as points on a 'scale of knowledge' or as a continuum (Henriksen, 1999; Laufer et al., 2004; Melka, 1997; Nation, 2001; Schmitt, 2010).

Griffin (1992) found that the two categories may overlap as the associations formed for productive or receptive learning are bi-directional, i.e. that receptive learning can result in productive knowledge and vice versa (as quoted in Nation, 2001:33). Considering this, productive knowledge could be understood as entailing receptive knowledge; some passive vocabulary may be very well known but never used and therefore never active. However, some researchers insist that seeing receptive/productive as a distinction is more useful than the notion of a scale/continuum. Meara (1997) believes that there is a threshold where a word that is known receptively becomes available for productive use. He found that a word could only be applied to productive use when at least one of the items the word was linked to in the mental lexicon is activated. The lexical item in question would become available for production only if that was the case. Schmitt (2010) continued this argument by saying that whether or not a word can be used productively is a matter of this word being linked to many other words, which can then act as retrieval routes for the word in question.

Several word-learning studies gave empirical evidence for differences between active and passive vocabulary knowledge. Waring (1997a) conducted a study on word retention and the results showed large individual differences between receptive and productive knowledge. The findings suggest that learners do not seem to be proficient at both learning types (Waring, 1997a; Rott, 1999; Webb, 2008). Min (2008) compared reading plus vocabulary enhancement activities and narrow reading and found differences when testing for receptive and productive knowledge. However, Read (2000) pointed out that different researchers made this distinction differently and that it therefore was far from clear what exactly defined receptive and productive knowledge. Results from studies investigating productive/receptive knowledge highly dependent on the types of tests that are used (Laufer & Goldstein, 2004). Besides, as different types of word knowledge are interrelated it is extremely difficult to design tests so that "answers to one do not affect the others" (Schmitt, 2010:80).

Therefore, instead of applying the receptive/passive - productive/active categorization, Laufer and Goldstein (2004; also Laufer et al., 2004) proposed to access a hierarchy of vocabulary skills ("modalities", Laufer et al., 2004:209) in order "to overcome the confusion between active and passive vocabulary". This hierarchy distinguished four levels of knowledge of meaning: 1) active recall (ability to supply the target word (hardest), 2) passive recall (ability to supply the meaning of a target word), 3) active recognition (ability to recognize the target word when given its meaning), and 4) passive recognition (the ability to recognize the meaning of a target word given meaning options - easiest). This approach was used in my study.

Linked to this, research has discerned several core factors of what makes words difficult to learn. These are how familiar a learner is with the word's pronunciation, spelling, morphological structure, grammatical and semantic properties, and learners' awareness of 'synforms' (i.e. words that can easily be confused with other words Laufer, 1988; 1989).

The receptive/productive distinction is also relevant in the way it specifies the breadth/depth distinction discussed above: possessing receptive/productive vocabulary knowledge "is essentially a question of having control of the item, of being able to access the word's form (production) and meaning (reception) in communication. Thus, receptive and productive knowledge is procedural rather than declarative in nature" (Jensen, 2005:78).

Finally, the terms 'passive' and 'active' are sometimes used instead of 'receptive' and 'productive' (e.g. Laufer, 1998). It should be noted, however, that 'passive' should not wrongly suggest that the learner is doing nothing. Due to this, this set of terms is sometimes seen as obsolete (e.g. Waring, 1997). In my study both set of terms 'receptive/productive' and 'passive/active' are used to express a finer distinction of vocabulary knowledge, which is necessary, for instance with reference to the different parts of the vocabulary test ('active' versus 'passive' recognition (Laufer & Goldstein, 2004; Laufer et al., 2004).

Clearly, this is no exhaustive discussion of what word knowledge is. Other distinctions were excluded. I decided to focus on those concepts which are most relevant to my study. It should have become clear that lexical competence is not simply a matter of knowing or not knowing a word. There are many unanswered questions and the intricacy of understanding lexical competences has consequences for how vocabulary develops. The following sections discuss what is known about this matter.

2.3. Acquiring Vocabulary Knowledge

Vocabulary acquisition is on the one hand seen as an endpoint (something has been learned/acquired), but on the other hand it is "viewed as a process beginning with input and culminating with integration of new linguistic information into an existing linguistic system" (Gass, 1999a:327). The latter perspective highlights the procedural character of the phenomenon and reflects the current view of research on vocabulary acquisition. While it is improbable that a new word can be acquired 'maturely'[8] in just one encounter, it is rather likely that word learning is a recursive, incremental process (Gass, 1999a). Accordingly, researchers attempting to model vocabulary acquisition present it as a multistage system in constant flux. Hatch and Brown (1995), for example, distinguished five stages that a word proceeds through from input to output. The model shows that there is a high number of words that never reach the final stage of competence, as they fail to proceed in the process. In a similar stage-based model, Jiang (2000) outlines how L1 knowledge both aids and interferes with L2 acquisition. Other models focussed on processing. Clark (1993) discerned three vocabulary learning processes: isolating word-forms in input, creating potential meanings, and mapping meaning onto form. Aitchison (1994) describes three lexical processes of labelling, packaging, and network building, and Haastrup and Henriksen (2000) also differentiate three phases: notice (detach from context), analyse (recognise links) and integrate (restructure network). The terms 'network building' and 'integrating' show the common presumption that a word does not have meaning in isolation; but rather derives it (at least partly) from interaction with other words in the language (Meara, 1996). These threefold models of vocabulary acquisition processes reflect Nation's categorization of form – meaning - use in his vocabulary knowledge model.

[8] 'Mature lexical entry': "Wesche (personal communication, August 1996) has suggested the term mature lexical entry to describe the type of mental representation that is beyond the initial knowledge gained through the process of fast mapping and reflects a more extended knowledge base of the kind an adult native speaker will have developed." Henriksen, 1999:311.

Henriksen (1999) discerned three dimensions of vocabulary competence (partial-precise, depth, receptive-productive), which can be understood as modelling vocabulary knowledge development, if word acquisition is seen as moving along these continua. The partial-precise knowledge development happens through mapping meaning onto form. Henriksen understands mapping as an incremental process of labelling and packaging, gradually "narrowing down [the word's] field of reference" (p. 312). Similarly, by developing a semantic network around a word through contextualised encounters, the dimension of depth is likely to develop from being superficial to become gradually more and more far-reaching. This concept of 'network building' can be linked to the notions underlying the learning from TIE. In this process, organisational links in the mental lexicon are strengthened, for instance through gradual differentiation within a certain lexical field and thus connected to related items at the paradigmatic and the syntagmatic level (Haastrup & Henriksen, 2000). It is a slow process and believed to continue throughout a person's life (Aitchison, 1994). In the process learners have to restructure their knowledge of individual words as well as their interrelations (Haastrup & Henriksen, 2000). At that point on the continuum, the learner's interlanguage system is modified and the new lexical item becomes part of an already existent network ("system" referring here to the lexical field, Ellis, 1994). These findings have put the notion of vocabulary learning into a new perspective because emphasis is placed on vocabulary acquisition as "system changing" (Henriksen, 1999:310). Hence, vocabulary acquisition is understood more as a matter of system learning than of item learning (Allendorff & Wode, 1981; Haastrup & Henriksen, 2000).

Providing learners with TIE, especially glossing, might be counterproductive to network building, i.e. promoting retention in long-term memory because it might deprive the reader of an opportunity to infer and activate ambient lexical information, thus mitigating processing. This makes sense as network building can be seen as parallel to the process of encoding, as both processes require a considerable mental effort (Haastrup & Henriksen, 2000b, Hulstijn, 1992). Other linguists agree that the more decisions a learner makes about a word, and the more cognitively demanding these decisions, the better the word is remembered (Laufer & Hulstijn, 2001; Morgan & Rinvolucri, 2004; Thornbury, 2002). If this is the case, TIE might be counterproductive to vocabulary acquisition. I hope that my study can contribute to a better understanding of the role TIE play in aiding or hindering such learning processes.

Henriksen's (1999) third dimension describes how vocabulary knowledge develops on a receptive-productive continuum. She acknowledges that it is not fully clear how this development is achieved. Factors such as automatization through repetition, and the quality of the network building process are likely to regulate this (Gass, 1988).

Schmitt (1998; 2010) worked on discriminating basic factors of vocabulary acquisition, too. He suggested that acquisition is likely to follow a developmental hierarchy, which means that specific word knowledge types are acquired before others, but that it is not fully clear how or in what order. His statement that vocabulary acquisition is incremental because not only meaning, but many other aspects need to be acquired, which most likely cannot all be processed at one encounter, has great relevance for TIE research. On the other hand, the incremental nature of the acquisition process is probable, because different knowledge facets themselves are acquired through several exposures, so that knowledge gradually develops along the continuum. Pronunciation, for example, is likely to be refined gradually through several exposures. This development includes moving back and forth along the continuum as forgetting "is a natural fact of learning" in the acquisition of both productive and receptive knowledge (Schmitt, 2010:23). Schmitt found empirical evidence for this dynamic in his 1998 study in which he tracked the development of different types of word knowledge.

Particularly relevant in the context of incidental learning from reading is Schmitt's suggestion that the first exposure to a given vocabulary item may already lead to the acquisition of "some sense of form and meaning" (p. 20), depending on the form of the exposure. If the word encountered in written form, for instance, it is likely that some salient part of the form rather than pronunciation is registered. Schmitt further assumes that certain types of word knowledge develop faster than others. He suggests word class as a criterion (1998). As mentioned above, other studies have found that among others factors such as parallels (cognates, Swan, 1997), particularly salient word forms or word length (Laufer, 1997b; 2013), play a role.

These models and theories illustrate schematically the highly complex phenomenon of vocabulary acquisition. To understand the underlying processes, it is necessary to have some knowledge of overarching principles of language learning and the cognitive processes involved. The principles underlying word learning from reading enhanced texts are introduced in the following sections. First, I focus on how notions such as attention and noticing are relevant for vocabulary learning.

2.3.1. Attention, Noticing And Awareness In Vocabulary Acquisition

The concept of attention has been regarded as "necessary in order to understand virtually every aspect of second language acquisition" (Schmidt, 2001:3). This seems particularly true in relation to TIEs. The hypothesized effects of input enhancement are based on two attention-related claims. The first claim is that the perceptual salience created by highlighting the input draws the learner's attention to the highlighted forms, and the second says that the learning of the attended form occurs based on the premise that attention is what mediates input and

intake (Izumi, 2002; Leow, 2015; Shook, 1999; VanPatten, 1990; see section 2.3.2). To approach this issue in the context of word learning, the Input Enhancement Hypothesis (Sharwood Smith 1991, 1993) explains that directing learners' attention to target words through TIEs would help them to first notice words and then potentially learn them and thus boost language learning (Boers et al., 2016; Leeman et al., 1995; Leow, 2000; Rott, 2007).
According to Schmidt (1990, 2001), attention can be defined as focusing on, as noticing, and as being aware of the items in the input. This characterisation touches upon three key terms in the field; attention, noticing and awareness. It seems probably impossible to separate attention and awareness completely and there is a common assumption that attention and awareness are two sides of the same coin. Moreover, opinions diverge with regards to which of the two mechanisms is more closely related to learning. Schmidt (1990, 1994, 1995) mostly placed greater importance on awareness. Schmidt (2001) put the emphasis on attention as the mechanism that controls access to awareness. According to him, attention must be regarded as a necessary precondition for any kind of learning; as he states: "The orthodox position in psychology is that there is little if any learning without attention" (2001:16). Robinson (1995, 2001) prefers to understand awareness as a by-product of focal attention. Others again (Tomlin & Villa, 1994; Truscott, 1998) claim that not awareness, but only input and attention are necessary for learning. Recently, SLA researchers have therefore called for more clarity with regards to which of these exactly is being measured (e.g. Godfroid et al., 2013; Godfroid & Schmidtke, 2013; Leow, 2015).

Han et al. (2008) highlighted that, according to information processing theory,

> "processing of information is (a) selective; (b) limited in that individuals can process two different types of information simultaneously and effectively only if the processing of one of the information types is automatized and requires little, if any, conscious attention; and (c) simultaneous processing of two different types of information that are not automatized can lead to inadequate processing of either or both types of information, i.e. to a 'trade-off' effect; and (d) is essential for action control and for learning" (Han et al., 2008:604; see also Leow, 2015; Schmidt, 1990; 2001; Skehan, 1996; VanPatten, 1996).

This trade-off effect, i.e. the question of potential simultaneous processing, is a pressing issue in the classroom language learning situation. Its implications for vocabulary acquisition are discussed throughout this thesis. It is desirable that vocabulary acquisition boosted by means of input enhancement should not be on the expense of text comprehension. Ideally, both, vocabulary acquisition and text comprehension should occur simultaneously.

Two other questions have dominated research into the role of attention for language learning. One is whether and how much attention is a necessary precondition for further processing and then acquisition. The other one is in how far internal processes can be directed and controlled. Both are relevant with respect to TIE and incidental/intentional learning of vocabulary from reading.

Regarding the first question, one of the most influential concepts is the Noticing Hypothesis (Schmidt, 1990, 1995, 2001). The major point is that allocation of attention resources is necessary for learning to take place, i.e. that noticing, but not necessarily rule understanding, is important for L2 acquisition (Schmidt, 2001). Noticing is seen as "a conscious registration of the contents of focal attention" (Jourdenais, 1995:186). A first encounter with a word may draw a learner's attention to that item, but this may not be enough. According to the Noticing Hypothesis, awareness (through attention) is necessary for learning. Schmidt (1995) posits two awareness levels: the 'noticing' level, the "conscious registration of the concurrence of some event" and awareness at the level of 'understanding', i.e. the "recognition of a general principle, rule or pattern" (p. 29). The latter is not required for input to be initially processed (Laufer & Hulstijn, 2001). Learners must pay attention to elements "on the surface structure of utterances in the input" in order to acquire them (Schmidt, 2001:5). What exactly these are, however, is not quite clear.

Subliminal language learning, Schmidt claims, is impossible (1990). He sees attention to input as a necessary precursor to hypothesis-formation and testing. According to Schmidt, attention is "what allows speakers to become aware of a mismatch or gap between what they can produce and what they need to produce" (Schmidt, 2001:6). In L2 vocabulary learning this means that words are not acquired unless they are consciously noticed and processed, i.e. that nothing can become *intake* without noticing (Han, et al., 2008). The latter, in particular is essential: processing of the input is necessary for learning to take place (Leow, 2015; see also section 2.3.2). Learners have to be actively involved with the L2 form, noticing alone is insufficient. Godfroid et al. (2013) examined to which extent L2 learners' level of attention to unknown words predicts word recognition ability. Taking into account the above mentioned "frequent disagreement among researchers about which cognitive process - attention or awareness - a given measure of noticing probes primarily" (p. 488), Godfroid et al. (2013) focused on noticing of unknown pseudo-words as *attention*. However, they did not disregard the role of awareness for learning. Rather, they hoped to shed light on whether teasing apart attention and awareness in the construct of noticing was useful: "If attention and awareness turn out to be flip sides of the same noticing coin, we would still need to determine how the two mechanisms should be mapped onto each other." (p. 485). The study's key finding was a direct, positive relationship between amount of (primarily focal) attention and amount of vocabulary learning. This supports Paribakht and Wesche's (1999) claim that "vocabulary

learning through reading is in some fundamental sense not 'incidental,' at least from the learner's perspective" (p. 215). In line with Laufer and Hulstijn's (2001) involvement load hyposthesis (see section 2.3.5), Godfroid et al. (2013) found that L2 learners were more likely to learn the words with which they engaged longer and, presumably, more deeply during reading. Godfroid et al.'s (2013) vocabulary learning study lends further evidence to Schmidt's noticing hypothesis "by showing that not only more awareness but also more attention lead to more learning (Schmidt, 1994, 1995, 2001)" (p. 509).

Other studies have specified that several factors, like the salience of the word, previous contact with the word, motivation, or "the realisation that the word fills a gap in their knowledge of the language" may be affecting noticing (Laufer & Hulstijn, 2001; Nation, 2001:63). For vocabulary learning it is assumed that noticing involves de-contextualisation, that the word is removed from its message context to be focused on as a language item. Noticing occurs when learners look up, deliberately study or have a word explained to them, or guess from context (Nation, 2001). Therefore, using TIEs, i.e. increasing the salience of L2 items in textual input, seems to be a good way of making reading processes more effective. There is plenty of evidence for the tenets of the Noticing Hypothesis (Allen et al., 1990; Doughty, 2003; Leow, 2000; 2015; Mackey & Philp, 1998; Rott, 2007; Schmidt, 1990; Schmidt and Frota; 1986) and the role of noticing for learning is now widely accepted. There are, however, also critical voices. Tomlin and Villa (1994) claimed that attention without awareness could lead to learning and that only one component of attention, detection, and not noticing was needed for further processing. According to their argument, awareness is not required for detection. Empirical evidence was also provided by Gass (1997) and Williams (2004, 2005). Carroll (1999) called the Noticing Hypothesis 'pre-theoretic', as it does not specify which properties of input are available for noticing and learning. Methodological criticism aims at the fact that the Noticing Hypothesis is not falsifiable, given that awareness cannot be measured precisely (Truscott, 1998). Another problem with noticing may be that even if forms are noticed, it is not clear as what they are noticed (e.g. as unknown and therefore necessary to learn?, As unknown and vital enough for text comprehension; Izumi, 2002.). Robinson (1995; 2001) coalesced Tomlin and Villa's view with the Noticing Hypothesis by applying a restricted understanding of 'noticing', which sees it as equivalent to Tomlin and Villa's 'detection within selective attention' concept (Schmidt, 2001), a view that Laufer and Hulstijn also follow (2001).

A stance on the matter might depend on the type of language item at hand. As not all linguistic elements are equally structured, they may differ in their communicative value, formal and functional complexity, semantic load, perceptual saliency (Cho, 2010; Han et al., 2008). Their noticeability therefore varies greatly. This is relevant within the context of TIE, because these inherent differences determine saliency and impact on the way different linguistic elements are learned. Therefore, input enhancement might not work the same way with all

language items, i.e. not all of them are equally processed and are not the same in terms of the effectiveness of instructional activities (Schmidt, 2001).

Even if details of the precise role and balance of awareness and attention in the Noticing Hypothesis still need to be fully explored, there seems to be widespread agreement on the importance of noticing for learning. As Robinson (2005:641) states, "Noticing certainly contributes to learning and retention, and […] consequently consciousness raising […], input enhancement […], processing instruction […], or focus on form […], which aim to induce it, are likely to be beneficial to learners."

Leow (2015) attempted to integrate several of the principal variables that have been identified by many empirical studies investigating the involved learning processes from exposure to L2 input and learners' eventual output (Gass, 1997; Tomlin & Vila, 1994; VanPatten, 1996): working memory, attention, awareness, depth of processing, and prior knowledge. He developed the 'Model of L2 learning in instructed SLA', which is an attempt to formulate a finer-grained model than previously existing versions (for instance, Schmidt's (1990) Noticing Hypothesis or Tomlin & Vila's (1994) Model of Input Processing in SLA). Leow's model is premised on the role of attention, i.e. the tenet that without minimal attention to input it is unlikely that L2 learning will occur. Leow situates the constructs of learning and awareness within an SLA theoretical framework and considers the dominant paradigms in cognitive psychology as well as in SLA studies. In order to describe the learning processes potentially leading from input, to intake, and to output, he discerns several different stages through which the learning process passes and positions attention and awareness therein. Leow (2015) bases his deliberations on the notion that language learning includes both internal processes (processing of input, intake, and knowledge) and products (intake and L2 knowledge/output as products). This is because, according to Leow, it is important "to differentiate between learning as a process, that is, an event taking place, and learning as a product, that is, something learned or internalized or produced" (pp. 21-22).

The first stages of Leow's (2015) model, 'input processing', seem most relevant for the study at hand, as the described processes are clearly linked to initial vocabulary learning from reading enhanced texts. The input processing at the first processing stage is largely dependent upon the level of attention the learner pays to input that is stored in working memory and which might be accompanied by depth of processing, cognitive registration, and level of awareness. In subsequently turning this input into (preliminary) intake, the second stage is divided into the three phases of attended intake (peripheral), detected intake (selective), or noticed intake (focal). According to Leow's model, it is possible for both detected and noticed intake, and to a substantially lesser extent, also peripherally attended intake, to be stored in

working memory and made available for subsequent recognition by L2 learners. However, they could also all be discarded if not minimally processed further (Godfroid et al., 2013).

Similarly, learning in the potentially following processing stages (intake and knowledge processing) depends on depth of processing, awareness levels and activation of old and new prior knowledge. Linguistic data processed with a low level of cognitive effort may most likely lead to implicit knowledge. On the other hand, processing with a higher level of awareness may facilitate the explicit integration of the intake into a learner's language knowledge system (see section 2.3.2 and 2.3.4). At the final processing stages, the ability to activate appropriate knowledge and monitoring one's own output or external feedback are additional contributing factors.

Besides the important role of attentional ressources, it is the significant role of a learner's processing capacity that particularly stands out in Leow's model. Leow posits that the role of awareness at different stages of the L2 learning process may be dependent upon the role played by depth of processing or how the L2 data are processed by the learner; i.e. a higher level of processing might lead to raised awareness levels, albeit not necessarily at the level of 'understanding' (Leow, 2001). Likewise, he further showed that intake (see section 2.3.2) may disappear from working memory if no further processing ensues.

The importance of attention distribution and noticing in the context of this study needs to be stressed again. These concepts are of great relevance because of their connection to task design and input enhancements. According to the 'cognition hypothesis' (Robinson, 2001; 2007) task complexity manipulations can directly affect L2 learning (Révész, 2009). Using TIEs can be one way of varying task complexity, e.g. when used to highten or lessen the amount of attention paid to decoding vocabulary. As attention is limited, and because any activity that draws upon it will interfere with other activities requiring it, language learning tasks must allocate attention strategically. This attention distribution is linked to the primary function of textual enhancement, namely, according to Han et al. (2008:601), "to serve as a priming device for learners' noticing of features in input, whose corollary may then be that what is noticed translates into acquisition". Although previous studies (Alanen, 1995; Cho, 2010; Doughty, 1991; Izumi, 2002; Jourdenais et al., 1995; Lee, 2007; Leow, 2001; Leow et al., 2003; Song, 2007; White, 1998; Wong, 2001, 2003) have yielded mixed results in terms of the effectiveness of input enhancement on noticing, many of these lend support to the claim that TIE draws learners' attention to form. The assumption that attentional resources must be allocated, and, in order for input enhancements to work, learners must attend to formal features in the input links TIE-use to the concept of 'Focus on Form', i.e. momentarily

focussing of attention on formal linguistic characteristics of language[9] (Long & Robinson, 1998; Long, 2000). Tavakoli (2013) claimed that "there is a relationship between input enhancement and Focus on Form, as both involve simultaneous attention to form and meaning in the input" (p. 172). There are parallels to the assumptions underlying TIE-use, as paying attention to form happens during tasks that are primarily meaning-based, i.e. occurs incidentally (Révész, 2009).

Thus, the concepts of attention, noticing, and awareness can play an important role concerning pedagogical practices (Leow, 2015), for instance regarding task design and task demands, because it is those that have been found to govern the attention paid to new words (Hulstijn, 2001; Keating, 2008; Schmidt, 2001). Task-based language learning research, such as Skehan and Foster's investigations of information processing (2001) and Robinson's (2005) study on cognitive complexity and task sequencing suggest that language learning tasks could be designed so that learners' opportunities to focus on form were increased, i.e. that modifying task complexity can affect language learning. For example, Révész (2009) investigated the link between task complexity, focus-on-form recasts and L2 morphosyntactic development and found that hightened task demands can have a positive effect on language development.

Although there is some research suggesting the possibility of unattended learning (see Leow, 2015 for an overview), this appears limited in scope and relevance for SLA. As Schmidt (2001:3) points out, "there is no doubt that attended learning is far superior, and for all practical purposes, attention is necessary for all aspects of L2 learning". In the context of classroom learning research, the relevance of this debate may be questionable. By following Baars' argument, I claim "that the important question is not whether there can be any learning without attention and conscious involvement but rather whether more attention results in more learning" (Baars, 1988 as quoted in Schmidt, 2001:30).

Another central question in this field is in how far attention can be directed or controlled (see above). In this context, the key terms are automaticity and controlled processing. Language processing control requires attention, is therefore slow, takes effort and is serial in nature (Levelt, 1989). It is, however, also controllable and can be modified by the learner, for instance in accordance with task demands. Controlled language processing is more typical of novice learners, while more accomplished learners typically feature more automaticity in their processing (Pellicer-Sánchez, 2012). Automatic processing, on the other hand, is often regarded as unconscious, effortless and fast (DeKeyser, 2001; Segalowitz & Hulstijn, 2005)

[9] As opposed to a pure focus on meaning in communication.

and as not sharing processing capacity with other ongoing processes (Levelt, 1989). Due to learners' limited processing capacity, the more that can be handled automatically, the more attentional resources are available for new information (Levelt, 1989; Sharwood Smith, 1993). In that, automaticity is vital for vocabulary acquisition.

Finally, in language learning, a necessary consequence of attention is encoding into memory. Only what subjects pay attention to is encoded in memory (Schmidt, 2001). The memory system is understood as consisting of different levels, the sensory store (records visual image), the short-term memory (converts image into meaningful information), and long-term memory. Knowledge is stored in the form of concepts or ideas (Alloway, 2005; Eskey, 2005; Juffs & Harrington, 2011; Williams, 1999). It is now believed that the short-term storage acts as 'working memory', a concept which refers to the structures and processes that humans use to temporarily store and control, but most significantly, to manipulate information in learning (Gass & Selinker, 2008; Leow, 2015; Miyake & Shah, 1999). Concerning TIE, it is important to remember that 'working memory' relates to the ability to juggle numerous language tasks. As this is a limited-capacity processing system, learners can only attend to a limited amount of material at a time. This puts constraints on how many words are processed, especially when texts are enhanced for word learning (Barcroft, 2002; Ellis, 2001; Rott, 2007; Van Patten, 1996). There is empirical evidence for the relationship between working memory and noticing and for a correlation between L1 and L2 working memory scores (Mackey et al., 2002; Miyake & Friedman, 1998). Furthermore, research linked phonological working memory capacity to the ease or efficiency with which novel words are learned (Papagno & Vallar, 1992; Service & Craik, 1993).

In conclusion, it seems that drawing attention to the input benefits processing, and that salience and meaningfulness of the input are major components "in determining the extent of that benefit" (Shook, 1999:67). These are key notions underlying the application of TIE in foreign language learning. However, ultimately, attention paid in the form of noticing is only the first step in the learning process. What has been noticed in the input needs to be transformed into intake. The following section reviews research that investigates the processes involved.

2.3.2. Information Processing: From Input-To-Intake

In relation to the processes involved in the success or failure of using TIE, another important distinction is the one between *input* and *intake*. Input enhancement may not succeed because the relevant (linguistic) structural properties of the utterance in the input have not been "registered by the processing system that is geared specifically to handle linguistic input […],

which means that the input has not become intake for acquisition" (Sharwood Smith, 1993:168). According to some researchers (Peters, 1985; Richards & Schmidt, 2002), input is the language that is received and extracted/segmented by a learner. Tomlin and Villa (1994) understand it as that part of the language that is detected by the learner. Intake, however, is that part of the input which the learner selects for further processing (Shook, 1999). Uptake refers to the processed items that are actually taken up for use (Milton, 2008). Leow (2015) sees output as the integral final stage of the learning process.

Related studies investigated which part of the language a learner is exposed to, is actually processed. Sharwood Smith (1993) reasoned that using the term input was actually misleading, as one could never know from "observation alone exactly what is processed by the learner at a given moment in time" and added that "samples of language may be overheard, marked as 'unknown', others may be processed." (p. 167). Furthermore, forms may be noticed perceptually, but not linguistically. Enhanced forms may attract attention, but may fall short of further processing (Leeman et al., 1995). Accordingly, "input enhancement implies that we can manipulate aspects of the input but make no further assumptions about the consequences of that input on the learner" (Sharwood Smith, 1993:176). Sharwood Smith used the term input in the sense of "potentially processible language data" which are made available, by chance or by design, to the language learner. Only the part of this input which has actually been processed by the learner and turned into knowledge of some kind is called intake (Sharwood Smith, 1993:167, Corder, 1967; Ferguson, 1971). Based on this understanding of input, Sharwood Smith considered it necessary to change the terminology in this debate. He introduced the term 'input enhancement' for attention-drawing activities, which were previously referred to as 'consciousness-raising'. The major difference between consciousness-raising and input enhancement is in reference to the input versus intake distinction: consciousness-raising implies that the learner's 'mental state' is altered by the input and no distinction is made between input and intake. In relation to input enhancement, however, it is assumed that while input can be manipulated, assumptions about success or failure of these changes for the learner are impossible.

Considering what happens with the input a learner encounters, i.e. how input is processed, research has taken either a comprehension or a use perspective (Jourdenais et al., 1995; Leeman et al., 1995) or focussed on foreign language reading processes (Leow, 1993; 1995a; 2015; Shook, 1994; 1999; VanPatten, 1996, 2000, 2002, 2007). Below, the latter strand is presented as it is most relevant for this study. The notion of 'attention distribution' is again important here. However, it is not quite clear how much attentional processing is necessary for input to become intake. There seems to be consensus that some attention is necessary,

but there are varied opinions about what type and amount (Jourdenais et al, 1995). Shook found that processing of (grammatical) items from input into intake required explicit attention, and that more meaningful items were processed before less meaningful items (Shook, 1994; 1999). Thus he confirmed VanPatten's findings that meaningfulness determines processing of input (VanPatten, 1996; 2002). VanPatten found that language learners rely on two processing strategies, firstly, on processing meaning. Only when meaning-processing has become automatic, conscious attention becomes available for processing form. Thus, conscious attention to form in the input competes with conscious attention to meaning. By extension, this means that only when input is easily understood can learners attend to form as part of the intake process.

Concepts such as attention and the input/intake distinction describe only one facet of vocabulary acquisition processes. Other learning conditions must be considered to fully understand the effects of TIE. As learners can be regarded as independent agents, it might be assumed that their intention to learn impacts language processing. This issue is discussed in the following section in the context of the incidental/intentional learning distinction.

2.3.3. Incidental And Intentional Learning

Traditionally, 'incidental' and 'intentional' learning were regarded as contrasting concepts. The distinction takes its origin from experimental psychology. The question of whether subjects were told about an upcoming test was used as a dividing line between 'incidental' and 'intentional' (Eyseneck, 1982, as cited in Laufer & Hulstijn, 2001; Zandieh & Jafarigohar, 2012). However, there is a wealth of other approaches to defining the one and the other form of learning. For instance, incidental learning is more generally understood as what takes place when "learners are focused on comprehending meaning rather than on the explicit goal of learning new words" (Wesche & Paribakht, 1999:176), so that it is a by-product of something else, usually comprehension (e.g. Gass, 1999; Laufer et al., 2004; Swanborn & de Glopper, 1999). Huckin and Coady used the term 'secondary learning', i.e. learning, which is not the target of the main cognitive activity (1999:182) or, as Gass put it, secondary to another pedagogical activity (1999a:320).

In the context of vocabulary learning, incidental learning was defined as learning that takes place when there is no conscious intention to do so (Barcroft, 2004; Hulstijn et al., 1996; Schmidt, 1994). Ellis (1994b) specified that this does not mean that the learner does not notice the word in question. It only means that attention is focussed on "understanding the passage as a whole, and memory for the new word comes as a natural result of this process, a conscious effort to learn being unnecessary" (p. 219). He discusses incidental learning in terms of learner attention, an approach which Gass finds problematic, as whether or not

something is the focus of deliberate attention is difficult to determine (Gass, 1999a). It is this link to elusive concepts such as noticing and attention that makes 'incidental' learning processes difficult to comprehend. Gass (1999a) found that "one of the difficulties in coming up with a good definition of incidental learning stems from the fact that there is no way to show that a given word was incidentally learned" (1999a:320; also Sharwood Smith, 1991; 1993). Accordingly, there have been doubts whether attention really is the crucial factor. Instead, Gass suggested that incidental vocabulary learning could be linked to, for instance, the nature of the word. She found that it is most likely to occur with cognates, when there is significant exposure, and when related L2 words are known (Gass, 1999a).

Taking a different approach, Hulstijn (2001) presented the view from cognitive psychology that "what is critical to lexical acquisition and retention is the nature of the input processing activities carried out by the learner, regardless of intention." (Hulstijn, 2001:268). For instance, he regards the quality and quantity of processing leading automaticity in accessing lexical features as more relevant. This line of argument was refined in the Involvement Load Hypothesis (see section 2.3.5). Essentially, it relates to the idea that "recognition of a *need* to learn (i.e. a gap) is the first step in actual learning. Whether that *need* is internally driven, […] or externally driven (e.g. by a teacher or textbook) is not important." (Gass, 1999:324).

While these approaches may be useful for a theoretical discussion about incidental/intentional learning, applied to real-life learning and research situations they might be insufficient. The methodological and educational meaning of this distinction seem to be two rather different phenomena. On a methodological level, the distinction is problematic, as repeated measure task-test-task-test designs, for example, inevitably risk to partly direct participants' attention to deliberate word learning rather than to meaning-oriented reading only. It seems to be accepted that some test scores may be due to intentional, rather than incidental learning. The incidental-intentional dichotomy has therefore been challenged (Barcroft, 2004; Paribakht & Wesche, 1997; Zimmerman, 1997). Barcroft suggested to conceptualize incidental/intentional vocabulary learning not as two dichotomous modes, but rather as orientations of learners while working on a task, situated on a continuum from intentional to incidental (Barcroft, 2004). The rationale behind the categories established on this continuum is that they require differing degrees of conscious processing. Barcroft claims that "a great deal of vocabulary learning may be neither purely incidental nor purely intentional […]. Different types of vocabulary learning can be viewed as points on a continuum between incidental and intentional because attention is not a dichotomous entity" (Barcroft, 2004:201).

The concept of such a learning continuum addresses problems of the incidental/intentional distinction that arise when looking at it from an educational perspective. This approach is in line with Paribakht and Wesche's (1999) argument that "vocabulary learning through reading

is in some fundamental sense not 'incidental', at least from the learner's perspective" (p. 215). Learners have their own intentions that belie the instructions they are given" (p. 176). They may

> "chose on their own to attempt to learn a word in varying degrees at any time; and it is difficult to determine the extent to which a given task may or may not invoke intentional learning even when learners have not been instructed to learn target words." (Barcroft, 2009b:87).

It seems important to consider that learners have their own agenda and that this fact can alter learning processes fundamentally. Gass (1999a) phrased this as the fact that students "are also their own teachers – they have their own focus of attention" (p. 321). Research revealed that learners predicted that there would be a test and therefore paid particular attention to certain words (Pellicer-Sánchez & Schmitt, 2010). Moreover, several other factors may contribute to whether or not a word is learned incidentally.

By suggesting that not only the distinction between incidental and intentional learning is problematic, but also the above-mentioned notion of the continuum, and that also the term 'incidental' was generally impracticable, Bruton et al. (2011) sparked a debate (see also Lopez, 2012; Reynolds, 2012). They suggest an alternative term to the incidental-intentional contrast, 'induced vocabulary salience', "which reflects an external intervention perspective, rather than an internal individual participant one" (Bruton et al., 2011:759). They relate their line of argument to that of Sharwood Smith (1991), who found that it was useful to talk of 'input enhancement' rather than of 'consciousness raising' (see above). Bruton et al. (2011) therefore argued that "for pedagogical L2 research purposes, the conclusion here is that it is more constructive to study external intervention, which might result in more or less (induced) attention on the part of the participants" (p. 764). This new direction of thinking and terminology is an ongoing discussion (Lopez, 2012; Reynolds, 2012).

From a practitioner's point of view, it seems useful to summarize this debate according Hulstijn's (2001) article on incidental/intentional vocabulary learning. He pointed out that 1) theoretically the distinction between the two learning concepts has become difficult to uphold, but that it was 2) methodologically essential for designing vocabulary learning experiments; and that "pedagogically, the distinction may have something to offer provided that teacher and learner are aware of points 1) and 2)" (p. 267).

There is some agreement on vocabulary learning situations. If student attention is directly engaged and focused on vocabulary, this offers a greater chance for vocabulary learning (Horst et al., 2005; Hulstijn, 1992, 2001; Laufer, 2005; Nation, 1990, 2001; Schmitt, 2008). Research investigating L2 incidental vocabulary learning has shown that vocabulary gains

tend to be lower and take longer (Day et al., 1991; Dupuy & Krashen, 1993; Horst, Cobb, & Meara, 1998; Hulstijn, 1992; Webb, 2005). However, as teaching all necessary words explicitly is impossible, incidental learning from extensive textual input is an important add-on to what can be achieved in a classroom.

In conclusion, three points in this debate seem particularly relevant for my study:

- To investigate the effect of 'induced vocabulary salience'; Bruton et al. (2011) suggest that empirical research should contrast the effect of single influences (e.g. enhanced versus unenhanced or different types of enhancements), examine the different combinations of influences (e.g., bolding plus marginal gloss versus bolding/no bolding plus linked gloss) or look at the effect of cumulative enhancements. This is what I did in my study.
- It should now be obvious that the dividing line between incidental and intentional learning is blurred, that these categories are somewhat arbitrary. Through task design one may lead to the other; task design may channel motivation and create a learning '*need'*. The potential effects of these aspects were investigated in my study, in particular in the interview data.
- Understanding incidental/intentional learning is difficult because the distinction between learner attention and pedagogically induced attention is crucial, but "these two may or may not fit" (Gass, 1999a:321). Here again, interview data collected in my study can shed light on the matter.

2.3.4. Explicit And Implicit Learning

Incidental/intentional learning can easily be confused with the explicit/implicit learning dichotomy (Dörnyei, 2009; Rieder, 2003). Sometimes the two seem to be conflated (Min, 2008; Webb & Nation, 2012). The confusion stems from the fact that these concepts overlap and that, just as with incidental/intentional, for distinguishing explicit/implicit learning, a learner's awareness of language and task are key issues. However, they "refer to different constructs in different domains of inquiry" (Hulstijn, 2012:2). While the incidental/intentional distinction takes a more motivational-methodological perspective on learning processes and focusses on intention, explicit versus implicit is concerned with the nature of the neuro-cognitive processing of language items[10] (Hulstijn, 2005; 2013). However, the explicit/implicit learning distinction is fervently debated and it is unlikely that all experts would agree on this point (Hulstijn, 2005; Rieder, 2003).

[10] Other possible confusions might occur with concepts such as explicit/implicit knowledge, declarative/procedural knowledge, explicit/implicit memory, attended/unattended learning, and explicit/implicit instruction. See Hulstijn, 2005; Muñoz, 2012; Rieder, 2003.

Ellis (1994a) defines implicit learning as “acquisition of knowledge about the underlying structure of a complex stimulus environment by a process which takes place naturally, simply or without conscious operations”. Explicit learning, however, “is a more conscious operation where the individual makes and tests hypotheses in a search for structure” (p. 1). It is also associated with selective, effortful processing, and the intent to “work out the concepts and rules with which these regularities can be captured” (Hulstijn, 2005:131). All the while implicit learning is “input processing without such an intention” and associated with automatic processing in the fashion typical of L1 acquisition (Hulstijn, 2005:131; Muñoz, 2012). However, learning is a dynamic process and the explicit/implicit distinction is not always clear (DeKeyser, 2003, Dörnyei, 2009). Ellis describes explicit/implicit as “dissociable but cooperative” concepts (2005:305) and it might be simplifying to see explicit/implicit learning as an either/or issue. Again, a continuum might be the more apt illustration (Muñoz, 2012; Reber, 2003), especially considering that learning leads to the fluid entity of ‘knowledge’ (Gass & Selinker, 2008).

So far, there is little empirical evidence for the efficacy of the explicit/implicit learning modes (DeKeyser, 2003). With respect to vocabulary acquisition, this type of research would be useful as it might show which words or parts of a word should be learned explicitly or implicitly. Ellis (1994c) claims that form, and perceptual and articulation aspects of new words are learned implicitly resulting from practice and frequent exposure. Meaning, and mapping meaning onto form, on the other hand, should rather be learned explicitly, as it requires conscious processing. However, these practical conclusions need to be treated with caution, as there is still uncertainty about the theoretical underpinnings of explicit/implicit learning. Hulstijn (2005) lists several factors, such as individual learner differences and the frequency and salience of the processed language features in the input as factors that affect the feasibility of one learning mode or the other.

This ongoing debate is linked to issues concerning input enhancements, in that it tries to pinpoint forms of processing that lead to effective learning, which TIE is concerned with. It is further relevant here as the explicit/implicit distinction has its origins in cognitive psychology, which in the early 1970s was dominated by the ‘depth of processing’ concept, which further shaped ideas important for understanding effects of TIE. This concept is described in the following section.

2.3.5. Levels Of Processing And Involvement Load

According to the ‘depth of processing hypothesis’, the mental effort made by learners when inferencing a word’s meaning aids the retention of words (Craik & Lockart, 1972; Hulstijn,

1992). The terms 'deep' and 'shallow' processing are rooted in memory research of cognitive psychology. The central idea of the hypothesis is that 'deeper' analysis of a stimulus leads to a more persistent memory trace, with 'depth' referring to a greater degree of semantic involvement. In other words, *richness* of encoding is decisive for learning and not simply the presence or absence of semantic coding. Craik and Lockhart (1972) claimed that "analysis proceeds through a series of sensory stages to levels associated with matching or pattern recognition and finally to semantic-associative stages of stimulus enrichment" (p. 675). It does not matter how much attention is devoted to form unless the quality of attention somehow changes to involve deeper and more elaborate processing (Craik, 2002). There are two types of processing, Type 1, where a subject merely repeats analyses that have already been carried out, and Type 2, where the processing of the stimulus continues on to a deeper level. Only the latter increases long-term retention (Craik & Lockhart, 1972; Stevick, 1976). According to Craik and Lockhart, what causes deeper processing are the usefulness to the subject, and amenability of the material to deeper processing; while the nature of the material, limited available processing capacity, and task demands may be sources of a failure to reach hinder deeper processing (Craik & Lockhart, 1972; Craik & Tulving, 1975).

As Craik and Lockhart failed to define what constitutes the different levels of processing and how those could be measured, their concept-framework was not accepted as a fully-fledged theory of cognitive processing (Baddeley, 1978). Still, it is still assumed that elaborateness of coding is of primary importance for learning, for example in the discussion about 'cognitive involvement load'.

This concept was formulated in the Involvement Load Hypothesis (ILH) and can be seen as an attempt to solve the problems criticized in Craik's and hypothesis, by refining the notion of 'depth of processing' and applying it to SLA. Laufer and Hulstijn (2001) presented categories that would make it possible to operationalise different levels of processing for incidental word learning in various types of tasks. They generated three categories, a motivational one, *need,* and two cognitive categories, *search* and *evaluation*. Together they constitute Involvement Load. Laufer and Hulstijn claimed that language learning tasks evoking a high Involvement Load lead to better learning: "Our basic assumption regarding vocabulary retention is that retention of hitherto unfamiliar words is conditional, in general, upon the degree of involvement in processing these words." (Laufer & Hulstijn, 2001:17).

One of the crucial questions Hulstijn and Laufer (2001) pose is whether "instructional tasks [can] be classified in terms of their vocabulary learning effectiveness" (p. 540). In accordance with Craik & Lockhart's research, they claim that retention of new words is promoted by "elaboration on features" (Laufer & Hulstijn, 2001:6) and quantity of associates (Hulstijn & Laufer, 2001:541), i.e. that the richness with which the material is encoded is critical.

They assumed that “word retention when processed incidentally is conditional upon the factors in a task” (p. 14). According to their hypothesis, such key factors are motivation (*need*) and cognition (*search*, *evaluation*), as they explain and predict learners’ success in the retention of unfamiliar words. Including a motivational, non-cognitive factor, *need*, seems vital as learning a language is influenced not only by cognitive processes, but integrated in a socio-cultural environment, in which emotions and personal dispositions play a decisive role. Laufer and Hulstijn understand *need* as the learner-internal *need* to achieve, which is primarily a creator of tension (Oxford & Shearin, 1994). This tension may be based on the drive to comply with task requirements (Skehan, 1989). It is activated when learners believe that a word is relevant to be learned. The goal was to attempt to operationalise the constructs of noticing, elaboration, or *need* at task-level so that tasks could be classified according to the degree of *need* they evoke in the learner.

Search and *evaluation* are the cognitive dimensions. Search is defined as “the attempt to find the meaning of an unknown L2 word or trying to find the L2 word form expressing a concept […] by consulting a dictionary or another authority” (Laufer & Hulstijn, 2001:14). *Evaluation* is the process setting in when a particular word has been chosen as a potential candidate for the slot in question. *Evaluation* is part of what Hulstijn called ‘elaborate processing’ (Hulstijn, 2001:270), i.e. retention is primarily determined by the nature of information processing.

Tasks can induce one, two or all three components for each word. As involvement load is defined as “the combination of the presence or absence of these three involvement factors” (Laufer & Hulstijn, 2001:15), and also determined by their “degrees of prominence” (Hulstijn & Laufer, 2001:544), this is what constitutes the depth/height/intensity of a task’s involvement load. A task with vocabulary help, for instance glosses, has a lower involvement load than a task without, as in the former no *search* or *evaluation* is needed, they claim. Thus “tasks differ in the involvement load they generate” (Laufer & Hulstijn, 2001:15) and teachers can manipulate a tasks’ involvement load according to the needs of a class. Several studies investigate, question, and provide empirical evidence for the claims of the Involvement Load Hypothesis (Eckerth & Tavakoli, 2012; Folse, 2006; Keating, 2008; Kim, 2008, Peters et al., 2009).

Given that the Involvement Load Hypothesis claims that a way has been found to tailor-make tasks, involvement load seems to be a very useful tool to categorize, asses and manipulate tasks for the benefit of the learners. However, in its present form the hypothesis has limitations. First, it is not fully clear what the three factors stand for. Laufer and Hulstijn acknowledge that more precise definitions and a more thorough theoretical link between the factors and theories of information processing are needed (Eckerth & Tavakoli, 2012; Huang et al., 2012; Keating,

2008; Kim, 2008). Secondly, the preannouncement of a subsequent vocabulary test in combination with a 10-minutes pre-task learning phase to review the target words, as done in Keating's, 2008 study, "move the learning conditions more to the 'intentional' end of the incidental-intentional learning orientation continuum" (Eckerth & Tavakoli, 2012:244) so that it is not clear what the results reveal about incidental vocabulary acquisition. Thirdly, Eckerth & Tavakoli (2012) suggest integrating participants' actual task behaviour and test attitude.

The Involvement Load Hypothesis has much relevance for my study. First, Laufer and Hulstijn use TIE as an instrument to implement the different involvement levels. Whether my study's results reflect these levels is considered in the Discussion. Secondly, Laufer and Hulstijn ask whether "the quality of exposure to new vocabulary during incidental encounters [can] compensate for the relatively limited amount of exposure which is characteristic of learning an L2 in a non-language speaking environment" (Laufer & Hulstijn, 2001:22). With this, they are addressing a classroom problem similar to what I am investigating in my study, whether the quality of an encounter with a novel word can be affected by the use of TIE.

The learning situation deserves consideration, as vocabulary acquisition is never a purely intellectual procedure, but always also a personal and social process. Therefore, to understand the effects of TIE, it is vital to consider how vocabulary is taught and learned in foreign language classrooms. The following sections present research exploring this.

2.4. Learning And Teaching L2 Vocabulary In The Classroom

In the following section, first the different approaches to vocabulary teaching are considered, while the second subsection presents research into vocabulary learning strategies. Thus, research representing the external and internal perspective on vocabulary learning and teaching in foreign language classrooms is reviewed.

2.4.1. Approaches To L2 Vocabulary Instruction

To describe the changing attitudes towards vocabulary teaching, I use the implicit/explicit distinction here in relation to fundamental differences in teaching styles. Definitions are concerned with "how directly and systematically" L2 components like grammar and vocabulary "need to be dealt with in language classrooms" (Dörnyej, 2009:269). Instruction is explicit/implicit when learners do/do not receive information concerning rules or instructions to attend to forms underlying the input, respectively (R. Ellis, 1994; Muñoz, 2012; Norris & Ortega, 2000). L1 acquisition, for example, is natural and mainly implicit. Research (e.g. Elley & Mangubhai, 1983; Waring & Takaki, 2003) has generally reached the conclusion that equally, implicit L2 vocabulary acquisition - albeit with small gains and after repeated exposure - is possible through reading.

For a long time, however, explicit instruction, with a specific focus on grammar, was believed to be the superior approach. Accordingly, the number of words introduced in foreign language classrooms was kept fairly low in order to not distract from learning what was regarded more important (Schmitt, 2000; Thornbury, 2002). Approaches like Audiolingualism or the Grammar-Translation-Method were "[n]ot concerned with developing productive L2 competence in the learners" (Dörnyei 2009:273), relied heavily on drilling, language analysis, error avoidance, and vocabulary was seen as negligible (Brown, 2000; Dörnyei, 2009; Harmer, 2001; Klapper, 2006; Schmitt, 2000).

In the early 1970s, a shift in emphasis became most salient in the form of the Communicative Language Teaching (Brown, 2000; Harmer, 2001). This approach stressed the significance of language functions, based on authenticity of language and materials (Klapper, 2006), i.e. Communicative Language Teaching is an altogether more acquisition-oriented approach. It underscores the importance of implicit learning situations as it follows the belief that language, also vocabulary, is learned through "learners' participatory experience in meaningful L2 interaction" (Dörnyei, 2009:276). The advent of Communicative Language Teaching caused "a major re-think of the role of vocabulary and vocabulary became an objective in its own right" (Thornbury, 2002:14).

Another influential approach to vocabulary teaching was Lewis' Lexical Approach, which is based on two interlinked findings. One is the assertion that "language does not consist of traditional grammar and vocabulary but often of multi-word prefabricated chunks" like lexical phrases, collocations, and idioms (Lewis 1997:3). The other, the realization that grammar and vocabulary are more "fundamentally linked" than imagined, became known through the rise of research in corpus linguistics (Schmitt, 2000:14). The ability to analyse large banks of language data stored in computers now allows sound statements about vocabulary (Harmer, 2001). This makes it easier to understand the "nature and functions of vocabulary, what it means to know a word, and how best to acquire vocabulary" (Morgan & Rinvolucri, 2004:3). Nowadays, in many foreign language classrooms the old-fashioned grammar-vocabulary distinction is avoided, because vocabulary and language learning are thought of in terms of "gradual mastery of underlying systems" rather than disconnected separate phenomena. Therefore, when new words are "introduced into the class it will be appropriate not simply to present and record the word but to explore the grammar of the word" (Lewis 1993:115). Beyond that, there is "no "right" or "best" way" as to the 'how' of vocabulary teaching. Decisions depend on the purpose and on the learners' individual dispositions (Schmitt, 2000:142).

One widely respected framework for designing a balanced vocabulary instruction is Nation's 'four strands' approach (Nation, 2001). His model of 'What is involved in knowing a word' (see

section 2.2.1) shows that to learn vocabulary effectively, a focus on developing both receptive and productive knowledge is necessary. Furthermore, it illustrates that an emphasis on learning a word's form, meaning, and use is vital. Accordingly, encountering a word only once is unlikely to be sufficient for acquisition in most senses. Nation's "four strands" offer a structure for teaching programmes that take account of this incremental nature of vocabulary acquisition. In a programme comprising the four strands of, meaning-focused input, language-focused (form-focused) learning, meaning-focused output, and fluency development, "vocabulary development will be optimal if there is a similar amount of time given to learning in each strand", and if learning occurs in all four strands (Webb & Nation, 2012:3).

The first strand focuses on learning "language items through listening and reading activities" (Nation, 2001:2). This, naturally, presupposes some vocabulary competence. The second strand, 'language-focused learning', is associated with the 'focus on form' principle which "aims to facilitate switches of attention from meaning to form during communication" Robinson, 2005:638). This is the strand with the clearest focus on direct teaching and learning of vocabulary and vocabulary learning strategies (VLS). Thirdly, there should be a strand in which learners develop their knowledge through speaking and writing activities. Finally, the established knowledge needs to be rehearsed to strengthen fluency, so that words can be "used or understood in a more native-like manner" (Webb & Nation, 2012:5). The value of this scheme lies in underlining the power of balance as a leading principle. For example the balance between explicit and implicit vocabulary work, between introducing, practicing and reviewing words. Purely task-based, meaning-focused teaching is substituted with instruction that includes explicit exercises. According to Sonbul (2012), "Nation (2001) provides convincing arguments why decontextualized form-focused activities might be useful in the L2 classroom". Furthermore, the "four strands" underline the immense importance of repetition and recycling of lexical items in different contexts (Joe, 1995), and what Schmitt described as the necessary "meaningful engagement with words over a number of recyclings" (2000:14). Thus, aspects like register or collocational links can be acquired.

However, teaching will only result in learning if students are ready and able to process the input they encounter. Effective processing depends for example on the availability of suitable language learning strategies.

2.4.2. Vocabulary Learning Strategies

The strategies that lead to effective vocabulary acquisition have been extensively researched (Cohen & Macaro, 2007; Cohen, 2011; Fraser, 1999; Gu, 2012; Kojic-Sabo & Lightbown, 1999; O'Malley & Chamot, 1990; Schmitt, 1997). Unfortunately, there is little consensus about both terminology and categorization of strategies.

By following a learner-centred argument mentioned in Schmitt (2010), in the context of this study I understand 'strategies' as purposeful efforts to "select, and then pursue, learning procedures that [learners] believe will increase their individual learning effectiveness" particularly to facilitate reading and vocabulary learning by individual learners (p. 91). As this study is primarily concerned with vocabulary learning from reading, I focus on related findings. However, an outline of the major research foci in the field is also relevant for understanding the strategic behaviour displayed in connection to TIE-use.

In addition to language proficiency, for which mainly significant positive correlations with strategy use were reported (Gu, 2003; Kojic-Sabo & Lightbown, 1999; Mizumoto & Takeuchi, 2008; Svensson, 2012), and the effectivity of different types of VLS (Fraser, 1999), vocabulary has been used as a learning outcome measure in this field (Barcroft, 2009; Gu & Johnson, 1996). Most research so far has demonstrated a meaningful relationship between VLS and results, but has overwhelmingly focussed on receptive knowledge. Furthermore, research has shown that there are no good or bad strategies, but that appropriateness and effectiveness, of VLS depend most of all on the learner - the learner's own purpose, stage of learning, preferences, and also which aspect of vocabulary is being focused on, and sociocultural supports or constraints (Gu, 2012; Schmitt, 1997). The following features of a good strategic vocabulary learner were found: flexibility, adaptability, creativity, and autonomy in strategy use. The skilful vocabulary learner is able to combine different strategies and relies on a wide strategy repertoire, which is also determined by personality, age, purpose for learning, and language type (Moir & Nation, 2002; Oxford & Nyikos, 1989). Findings show that successful strategy users predominantly apply deep processing VLS (e.g. inferencing) rather than mechanical ones (e.g. repetition; Gu & Johnson, 1996; Lawson & Hogben, 1996; O'Malley & Chamot, 1990; Schmitt, 1997), have the ability to self-regulate their personal strategy use (Tseng et al., 2006), and are able to analyse a task for its strategic demands (Gu, 2012).

These findings suggest that one key issue in school practice must be "learner strategy training", which raises students' awareness for learning strategies and individual learning styles (Mochizuki 1999:102). The results of Schmitt's (1997) study indicate that learners may be willing to try new strategies if they are introduced to and instructed in them. Research has shown that strategies are teachable (Jooneghani et al., 2012; Schmitt, 2000).

Another important concern in strategy research has been the classification of the many different types of learner actions that may be described as 'strategies'. O'Malley & Chamot (1990) differentiated three types of general language learning strategies, 'metacognitive', 'cognitive' and 'social/affective'; and Oxford's (1990) Strategy Inventory for Language Learning classification system separated out a fourth category; 'memory strategies'. To

categorise the available knowledge particularly regarding vocabulary learning strategies, Schmitt (1997) compiled a taxonomy for individual VLS; using Oxford's four VLS groups he added 'determination strategies' as a fifth group to cover for situations where meanings of new words are discovered without recourse to help from other people. Following Nation (1990), Schmitt also integrated the 'discovery' (discovery of word meanings) – 'consolidation' (remembering word meanings) distinction into his taxonomy. Thus, he assembled six categories (discovery: social, determination and consolidating: social, memory, cognitive and metacognitive) categorising almost 60 different VLS. Due to its elaborateness, I applied Schmitt's taxonomy as a template and adapted it to categorize the strategies that emerged in my data.

In addition to knowing about VLS, knowledge of reading strategy use is paramount for understanding TIE-related research, as their use directly influences how well learners perform in language tasks. Bowles (2004), for example, found in the analysis of think-aloud data that participants had applied two distinct reading strategies to reading identical material (reading through the L2 text without stopping for comprehension, versus reading while translating into the L1), and that this had led to very different results in her vocabulary tests (noticing versus not noticing of target words). Other research found that successful L2 readers had a 'unitary view of language', i.e. they were aware of similarities and differences between languages and make "use of that knowledge in employing strategies for reading comprehension in their L2. Less successful L2 readers paid more attention to getting the task done than to comprehension" (Jimenez et al., 1996 in Erler 2007:195f).

Research on how reading links to vocabulary acquisition is reviewed in the following section.

2.5. Reading And Vocabulary Learning

Reading has the potential to expose learners to meaning-bearing, semantically, syntactically, and pragmatically rich input that is vital for developing language in general and the skill of reading itself. It is therefore indispensable for vocabulary acquisition (Krashen, 1993; Pellicer-Sánchez, 2012b; Rott, 2007; Wesche & Paribakht, 1999). It seems that the relationship between reading and vocabulary is a reciprocal and circular one: Eskey (2005) found that an excellent way to "acquire the extensive vocabulary required for reading widely in a second language is reading itself, and […] a prerequisite for such reading is an extensive vocabulary" (p. 567).

The following sections present research in the field of L2 reading and findings regarding the importance of reading for vocabulary learning.

2.5.1. Reading In A Foreign Language

Reading was originally understood as a 'bottom-up' decoding process, i.e. a step-by-step sequential 'text to brain' approach, where a reader's focus moves from understanding part of a text to the whole text (Eskey, 2005). Later, it was understood as a 'top-down' comprehension process, in which readers approach whole chunks of a text according to pre-formed conceptions of meaning, and where prediction and confirmation prevail. This model emphasises what the reader brings to the text (Gettys, 2001). Research on schema theory, which investigates how humans organize information in bundled knowledge structures or 'schemata', was helpful to understand how readers store and structure information in networks of related ideas (Eskey, 2005; Gettys, 2001; VanPatten, 1996, 2002). Nowadays, however, reading is seen as "an active, purposeful, and creative mental process" and interactive models of reading, describing a balanced combination of top-down and bottom-up processes, dominate (Eskey, 2005:564). It is also believed, that reading involves not only cognitive, but also social factors.

The overarching objective of L2 reading research is to characterize successful L2 reading. It confirmed that an effective reader needs a certain mastery of the language itself. As explained below, it was found that several other factors also play a role for reading success.

Reading has been linked to language proficiency. For non-native readers, proficiency is the decisive factor, as reading begins with decoding of language. According to Clarke's "short-circuit" hypothesis, readers need to have a certain level of language proficiency in order to be able to transfer L1 reading skills to reading texts in the L2" (Clarke, 1980; as cited in Eskey, 2005:566; Hulstijn, 2001; Koda, 2005). This competency threshold varies individually and is affected by motivation and knowledge. This degree cannot be compensated for by knowledge of reading goals, text characteristics and reading strategies.

The proficiency level determines the ability to master different levels of L2 text difficulty. Research explored what percentage of L2 words in a text needs to be known to facilitate comprehension. It was suggested that for sufficient 'lexical coverage' in L2 reading knowledge of 98% of a text's words is needed (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Laufer & Yano, 2001; Pellicer-Sánchez, 2013). This requires familiarity with up to 9000 word families (Nation, 2006). However, other research suggested not to understand this percentage as a one-size-fits-all number. Schmitt, Jiang & Grabe (2011) proposed the understanding of lexical coverage as a linear progression, where larger coverages lead to better

comprehension. Moreover, text coverage is not the only matter that determines text difficulty. The ideas a text conveys, the background knowledge that is needed to understand these, and the learner's awareness of how these ideas are structured are also influential factors (Grabe, 2009; Riley, 1993; Rott, 2007).

Automaticity is also closely linked to language proficiency. Efficient reading builds on largely automatic processes such as scanning and skimming (Carver, 1990; Hulstijn, 2001). Lack of automaticity can be problematic with respect to vocabulary acquisition and effective reading. A causal relationship between word recognition efficiency and reading comprehension was found for both L1 and L2 learning (Koda, 1996). If an L2 reader has not yet passed the necessary competency threshold, automatic word recognition is not guaranteed and the learner is likely to become frustrated. Therefore, "care must be taken not to immerse readers in texts that are lexically beyond them, which does in fact reduce reading to a kind of guessing game" (Eskey, 2005:567). However, the importance of using contextual clues for vocabulary learning and text comprehension has been validated by extensive research (e.g. Fukkink et al., 2001; Li, 1988; Mori, 2002; Svensson, 2012). While guessing from context and inferencing are naturally occurring reading strategies, they might not be effective for decoding, because they entail the risk of misinterpretations, i.e. context offers different types of clues, non-clues and misleading clues that the L2 learner has to discern from useful clues (Folse, 2006; Laufer, 2003). Hulstijn (1992) found that using natural context to guess word meanings is a very complex and error-prone process for L2 learners. Moreover, successful inferencing requires a large vocabulary in order to interpret context clues (Folse, 2012). Higher proficiency leads to more effective guessing and puts lower-proficiency learners at a disadvantage (Folse, 2006). Rather, what beginning L2 readers need in order for automaticity to develop is time to acquire the language through different means of exposure so that they can slowly accumulate experience with L2 reading (Eskey, 2005).

To bring about the needed competency and automaticity/fluency, L2 classroom learners are often confronted with two approaches, extensive and intensive reading. Extensive reading requires L2 students to read substantial amounts of L2 texts, often independently. Book flood programmes are a typical example[11] (Nation & Wang, 1999). There is evidence demonstrating that reading extensively, when done consistently over a long period of time, leads to better reading comprehension as well as improved abilities in several other language areas (Grabe, 2009a). According to research on cognitive processing and learning theory, there is strong evidence that extensive reading does support vocabulary growth (Grabe, 2009b). Despite this,

[11] See the seminal 'Fiji book flood study' (Elley & Mangubhai, 1981).

the role of extensive reading in classrooms around the world is small (Laufer, 2003). This is, among other things, due to time constraints. A more prevalent variation is 'narrow reading', the reading of numerous authentic texts which all regard the same topic (Schmitt & Carter, 2000).

One concern raised about extensive reading programs regards the authenticity of texts and whether or not it is a requirement for student reading (Grabe, 2009b). Some studies suggest that authenticity can affect the success of L2 reading, as it is naturally linked to text difficulty (Lyman-Hager & Davis, 1996; O'Donnell 2012). Lyman-Hager and Davis (1996) found that readers of authentic texts benefitted significantly more from enhanced texts than those working with simplified texts. It is, however, not always easy to determine what an authentic text is, and not always feasible to use authentic materials classroom settings.

In contrast to extensive reading, in intensive reading programmes learners work with small amounts of text "to make various points about the nature of texts and the reading process" (Eskey, 2005:574). Often, these activities rely on comprehension questions. It can be assumed that intensive reading is the type of reading experience common for most L2 learners (Macalister, 2011). It usually includes teacher-supported reading for comprehension with readers' attention drawn to certain words via textual (or other) input enhancement for the purpose of vocabulary acquisition. Intensive reading activities can be seen as being situated in the language-focussed or (Nation, 2009) meaning-focussed input strand (Macalister, 2011). A criticism of intensive reading is that while this type of reading activity does *practise*, it does not in fact *teach* reading or skills development (Nation, 2006).

2.5.2. Vocabulary Learning Through Reading

Based on his Input Hypothesis, i.e. that successful L2 learning results from "comprehensible input as the essential external ingredient" (Coady, 1997:225), Krashen (1989) suggested that L2 learning works the same way as L1 learning and that vocabulary and spelling are most efficiently acquired by receiving comprehensible input while reading. However, his analysis was overwhelmingly based on native-speaker studies rather than on L2 learners.

Setting the groundwork for modern research on L1 vocabulary learning from reading, Nagy et al. (1985) empirically substantiated the long held claim that most L1 words are learned from context and proposed extensive reading to foster L1 vocabulary growth. Swanborn and DeGlopper (1999) conducted a statistical meta-analysis of 20 incidental L1 word-learning-from-reading studies. They concluded that students learn about 15% of the unknown L1 words they encounter when reading. This number promises substantial word gains from L1 reading.

Findings from such studies raised the question of whether the same could be assumed for L2 vocabulary acquisition through reading. Scholarly views on what role reading can play for incidental L2 vocabulary learning vary enormously.

Given the huge amount of words that are needed to read independently - Nation (2006) assumes familiarity with ca. 8,000-9,000 word families as necessary - it seems obvious that explicit instruction cannot be sufficient to reach this goal. In particular, the 'mid-frequency vocabulary', i.e. words at the 3,000-9,000 levels (Nation, 2006) is most problematic. It is mostly not considered important enough to be taught directly and not rare enough to be ignored, but it is still necessary to communicate successfully (Pellicer-Sánchez & Schmitt, 2010). Due to time constraints, in classrooms it receives little attention. It is also problematic because it is very scarce in most texts. Nation found that only 4.9 % of the tokens in the BNC came from the combined 4,000-9,000 frequency levels, although this made up 25.25% of the types (Nation, 2006). Accordingly, while intentional learning and explicit teaching may be the most effective approach for the learning of many words, incidental learning is indispensable (Laufer, 2005; Schmitt, 2008).

Saragi et al. (1978) conducted a study in which incidental L1 vocabulary learning from reading was the dominant approach. Their participants read the novel A Clockwork Orange, which contains the hip-talk language 'nadsat' and were tested on the incidental acquisition of nadsat words. Their subjects correctly identified the meaning of 75% of the target words. Horst et al. (1998) replicated this study with L2 learners. On average, participants acquired five new words from a book of 21,000 words. This equates gains in word meaning recognition of 22% of the target words. Lahav (1996) reported an average learning rate of 3-4 words from four simplified readers with 20,000 words each. The tests in Brown et al.'s (2008) study of vocabulary learning from reading and listening showed gains and long-term retention of word form-meaning recognition from incidental vocabulary learning. Waring & Takaki (2003) investigated the vocabulary acquisition of Japanese learners' of English from reading a graded reader. Their findings suggest that very little new vocabulary is retained from reading graded readers and that incidental vocabulary learning from reading aids recognition more than production of lexical knowledge. Concerning the effect of retention over time it was shown that recall rates dropped more quickly than those for recognition. Their conclusion was that reading graded readers is not very effective for building new vocabulary, but is better suited for manifesting and enriching already known vocabulary. All these outcomes suggest that the important question for research is not whether vocabulary learning from L2 reading is possible, but rather whether it is effective.

Vocabulary uptake from reading has generally been found to be small and an error prone process (Horst et al., 1998; Nation, 2001; Pellicer-Sánchez & Schmitt, 2010; Peters et al., 2009, Schmitt, 2000). This fact led researchers in the field to challenge the position that L2 reading inevitably has a decisive impact on lexical growth and retention (Laufer, 2003, 2005; Min, 2008). According to Laufer's (2003) study in which she compared reading with word-focused tasks, four assumptions are prerequisite for the claim that vocabulary is primarily learned through reading. One is that unfamiliar words have to be noticed to be learned, the second concerns learners' guessing ability, that they can use contextual clues successfully, particularly if the word is deemed relevant for comprehension. The third is that guessing may lead to retention. Finally, it is assumed that frequent encounters with words in context heighten the probability that the word will be retained.

Laufer criticises these assumptions. Concerning noticing, she points out that L2 learners do not necessarily recognize words as being unfamiliar[12]. Therefore, she concludes that noticing cannot be taken for granted. With respect to guessing from context, she emphasizes that comprehension is not equivalent to retention. Especially with L2 learners, guessing from context may be unsuccessful. Furthermore, Laufer highlights that noticing only happens when learners find the unknown word important, otherwise it might be ignored. Moreover, guessing only leads to retention when some processing effort is made. This, however, interrupts the reading flow, which is why it is likely that learners will not bother to stop and think. Finally, concerning the cumulative-gain assumption, she points out that the number of encounters needed for a word to be acquired is largely unknown. Furthermore, it takes perfect reading conditions, encountering vast number of words and extensive amounts of time. These opportunities are usually restricted in a normal classroom situation, Laufer concludes that the frequency argument cannot be granted either (Laufer, 2003).

However, frequency has been shown to be a decisive factor for effective vocabulary acquisition. For example, Pellicer-Sánchez and Schmitt conclude in their review of vocabulary-learning-from-reading studies that "L2 learners can make meaningful lexical gains from reading, if they get enough exposures" (2010:35). This seems especially necessary for deeper learning to take place (Schmitt, 2000). For L2 contexts, at least 10-12 exposures have been proposed necessary (Horst et al., 2005; Waring & Takaki, 2003; Zahar et al., 2001). Still, Laufer's study makes an important point. It showed that "if a word is practised in a productive word-focused task, its meaning has a better chance to be remembered than if the word is encountered in text, even when it is noticed and looked up in a dictionary" (Laufer, 2003:581). Her conclusion was, therefore, that in instructed L2 contexts, reading only is unlikely to be the

[12] See the method chapter for a discussion of phenomena such as 'synforms' and 'deceptive transparency'.

most effective way to acquire vocabulary. However, the fact that most of Laufer's (2003) experiments were based on a single encounter of unknown words in texts is problematic, rendering the results less applicable to authentic instructional contexts (Min, 2009). Many studies have confirmed that reading-plus approaches are usually more successful than reading-only and it seems to be the common answer to the question of how reading can make a contribution to effective L2 vocabulary acquisition (Laufer, 2003; Paribakht & Wesche 1997, 1998, 1999).

Peters et al. (2009), in a study investigating the effect of test announcement, support this. They found that successful L2 vocabulary acquisition through reading is contingent on three factors. First, learners discover meaning of unknown words; secondly, the lexical information is processed elaborately; and, thirdly, form-meaning relationships of the words are reinforced through repetition (2009). The second point was also confirmed by Haastrup (1991). She found that words which are guessed with some difficulty will be remembered better since difficulty leads to processing effort, which in turn, creates a more distinctive memory trace. In a study investigating word learning from reading an authentic novel, Pellicer-Sánchez and Schmitt (2010) adapted the 'Clockwork Orange' methodology (Saragi et al., 1978, see above) and found that incidental vocabulary learning occurred from reading, i.e. learning was measurable for 9.39 out of 34 target words (Pellicer-Sánchez & Schmitt, 2010). The highest scores were measured for meaning recognition, but very low at recall level. Mastery at recall level is very difficult to achieve from only incidental receptive learning. Their results match previous findings (Schmitt, 2008). Overall, they confirmed the findings from previous studies that vocabulary learning from incidental reading is possible but ineffective and "the amount of incidental learning is unlikely to match the amount available from an explicit teaching approach" (Pellicer-Sánchez & Schmitt, 2010:43). Vocabulary instruction should involve a combination of both incidental and intentional forms of vocabulary learning.

However, given the time constraints in classrooms, it is often not feasible to link reading and explicit vocabulary training. Frequently, practitioners and textbooks use a compromise and instead enhance important vocabulary in intensive reading materials. The assumption is that both comprehension and vocabulary acquisition might be aided in this way. Several studies have investigated the effect of such textual input enhancements to see if a heightened quality of input would result in higher vocabulary learning gains. This field of research is presented in the following sections.

2.6 Effects Of Textual Input Enhancement

To outline the nature of input enhancements, three important dichotomies have to be explained. First, Sharwood Smith (1991; 1993) distinguished between *external* and *internal* input enhancement. External input salience, for instance created by a teacher, is distinct from input enhanced by an insider, i.e. the learners themselves (see also Han et al., 2008). This is because "externally generated input enhancement does not automatically imply the internalization of that enhancement by the learner" (Sharwood Smith, 1991:131). As is discussed later, this potential lack of internalization may impinge on the efficacy of the input enhancement. Secondly, different types of induced salience are *positive* versus *negative*. Positive enhancements make certain correct forms in the input more salient, while negative input enhancements flags up given forms as incorrect. Thirdly, input enhancement is also differentiated considering the degree of elaboration it initiates. According to Sharwood Smith (1993), boldfacing the target forms, for example, would be an unelaborated form of salience, with no appeal to metalinguistic knowledge, while, on the other hand, explaining a (grammatical) construction using metalinguistic terminology would be a highly elaborate form of enhancing the input. White (1998) proposed a differentiation between more or less explicit enhancement forms when pointing out that, for example, textual enhancements need to be considered as being more explicit than input flooding, but less explicit than rule explanation.

Working with TIEs means putting deliberate focus on certain properties "of language with the purpose of facilitating the development of L2 knowledge" (Sharwood Smith, 1991:118). The primary question for researchers has been to find out whether this facilitation succeeds and under which circumstances. Previous research revealed a mixed picture and some conflicting findings. This is mainly due to diverging methodologies, differences in the degree of text authenticity, text difficulty, the number of enhancements in the texts, or learners' proficiency level. Overall, however, research supports the view that TIEs aid vocabulary acquisition.

A significant number of studies were concerned with the effect of enhancements on noticing and acquisition of grammatical forms. Often TIEs were compared to other attention-raising strategies such as rule teaching. Several studies reported positive effects of TIEs (Alanen, 1995; Cho, 2010; Jourdenais et al., 1995; Lee, 2007; Russell, 2012; Shook, 1994; White, 1998; Winke 2013). They showed that TIEs did not only facilitate the noticing of target forms, but actually contributed to a heightened acquisition of new grammar forms. They also showed that combining different forms of TIEs was especially effective (e.g. Jourdenais et al., 1995; Shook, 2004; Wong, 2001). Other studies, however, found that the enhancements had only a limited impact (Izumi, 2002; Leow, 1997; 2001; Leow et al., 2003; Overstreet, 1998; Wong,

2003). In these studies, it was shown that, while using TIEs had a positive effect, other forms of input, for instance explicit grammar teaching (Alanen, 1995; Izumi, 2002) or input flood (White, 1998), were significantly more effective. On the other hand, some studies found no or negative effects of TIE on the acquisition of grammatical forms (Leow, 1997, 2001, 2003). Others reported mixed results, in that they testified negative effects on text comprehension or no significant impact (e.g. Overstreet, 1998). It is difficult to draw any generalizable conclusions from these findings, as these researchers worked with different types of enhancements, for different purposes, and used different testing instruments. Therefore, these findings are not reported in more detail. However, some of these grammar-focussed studies informed my research questions, particularly in relation to the effectiveness of particular forms of TIE, and are therefore included here.

In addition to research investigating how TIE can further grammar learning, there is a considerable number of studies that examine whether and how TIEs aid vocabulary acquisition and reading comprehension. Particularly relevant for my research are studies that investigate the effectiveness of different types of TIE, for instance typographical enhancements or glosses, and studies examining reader preferences, attitudes towards, and learners' way of working with input enhancements.

TIE have also been investigated with respect to other aspects, for instance advantages and disadvantages of different locations of the enhancements (Holley & King, 1971; Jacobs et al., 1994; O'Donnell, 2012; Stewart & Cross, 1993); or the effects of different types of presentation i.e., computerized versus print medium (Bowles, 2004; Gascoigne, 2006; Lomicka, 1998; Rashtchi & Aghili, 2014). Moreover, unanswered questions regarding how TIEs are used for aiding vocabulary acquisition and reading comprehension are introduced. This literature review concludes with the research questions, which result from questions which have been left unanswered in previous research.

Positive effects of TIEs on word learning and reading comprehension were found in several studies (e.g. Bowles, 2004, Hulstijn et al., 1996; Ko, 2005; Rott, 2002; Tabatabei, 2011). However, some reveal that even purposefully manipulated input does not necessarily result in higher language intake than without such changes (e.g. Holley & King, 1971; Jacobs et al., 1994; Leow, 2001; Rott, 2007). Overstreet (1998) and Lee (2007) found negative effects of enhancements on L2 reading comprehension. Reasons for success and failure of TIEs were investigated and various conditions were discerned as necessary for TIEs to lead to learning. Over the years, the focus of TIE-research has shifted from only investigating whether there are effects at all to examining which enhancement types are most effective. In the following sections, studies concerned with all of these aspects are reviewed.

2.6.1. Effects Of Different TIEs Types On Vocabulary Acquisition And Reading Comprehension

Research in this field has been concerned with the question of which types of TIE are most effective and for which purposes. The findings revealed the following twelve aspects to be of particular importance: learner proficiency, glossing for vocabulary acquisition may be distracting from comprehension, task elaboration, level of obtrusiveness of the TIEs, whether TIEs are of simple, combined[13], or compound[14] form, prior topic knowledge, which language is used for glossing, whether TIEs are internal or external, frequency as a form of enhancement, why TIEs might not work, and their effects on the acquisition of collocations and on long-term retention. Research investigating these issues is presented in the following subsections.

Proficiency

For TIEs to be used effectively, the reading material needs to have an appropriate difficulty level. Therefore, several researchers included proficiency in their investigations. Cheng and Good (2009) investigated the effects of three gloss types (L1 glosses with L2 examples, L1 in-text-glosses, and L1 glosses placed below the text) on vocabulary production and comprehension. Proficiency was included as one variable, and forgetting patterns were analysed to compare the results from immediate and delayed recall tests. In the data collected from 135 participants, no significant effects were found for reading comprehension. Glossing emerged to have a generally positive effect on vocabulary retention, but the effect was significant for specific proficiency levels only (intermediate, and higher-intermediate learners). The most effective type of gloss were the L1 glosses with L2 examples, followed by in-text L1 glosses. Data collected in the second delayed post-test showed a slight increase in vocabulary recall. Cheng and Good assumed that this increase could be credited to the repeated review of the glosses. Their study prompted a comment by Taylor (2010), who remarked that for some proficiency levels (beginners and advanced learners), glossing had “the potential of getting in the way of reading comprehension” and that text difficulty was decisive (Taylor, 2010:353). The impact of proficiency on success in using glosses was also investigated in other studies (e.g. Bland et al., 1990; Jacobs et al., 1994; Ko, 2005, O’Donnell, 2012). On the one hand, these studies suggest that the approach to gloss usage is linked to readers’ proficiency levels: Less proficient learners use glosses for different purposes and in different ways than more

[13] Enhancement forms that combine different types of enhancements, e.g. bold-printing and glossing.
[14] Enhancements that are combined with other salience-increasing exercises.

proficient learners. It seems that learners have to overcome a general language proficiency threshold to be able to use TIEs the way advanced learners do. On the other hand, different proficiency levels seemingly also relate to different gloss types, for instance more proficient learners use and benefit more from using L2 glosses, than less proficient learners who rely more on L1 glosses (Cheng & Good, 2009; Ko, 2005). Therefore, for investigating the effectiveness of TIE and reading comprehension, it is crucial to consider both proficiency levels of research participants and to find texts at appropriate difficulty levels.

Processing Vocabulary versus Reading Comprehension?

In his response to Cheng and Good, Taylor (2010) suggests that glossing may distract learners from processing reading comprehension. This assumption is the second of the factors, listed above, that have become prominent in TIE research. According to my experience as a foreign languages teacher, in real-life learning situations, TIEs are used for facilitating vocabulary acquisition, comprehension, or often, both. In many cases, the purpose of using input enhancements for one end or the other is not clearly distinguished. Even though there is research exploring the effects of TIEs on either exclusively vocabulary acquisition (Al-Seghayer, 2005; Holley & King, 1971; Hulstijn et al., 1996; Kim, 2006; Min, 2008, Rashtchi & Aghili, 2014; Xu, 2010; Yoshii, 2006) or reading comprehension (Bell & LeBlanc, 2000; Brantmeier et al., 2012; Davis, 1989; Fahimipour & Hashemian, 2013; Ko, 2005; Lomicka, 1998; O'Donnell, 2012; Roby, 1998), findings and methodologies suggest that the two concepts somehow overlap (O'Donnell, 2012; Taylor, 2010) and have to be considered together. As Han et al. (2008) state, TIEs are "premised on the notion that learners must comprehend what they read or hear before their attention can be drawn to form within the input" (p. 599). Therefore, it seems appropriate to avoid a distinction of these aspects.

Some studies found a positive relationship between enhancement use and comprehension (Bell & LeBlanc, 2000; Holley & King, 1971, Watanabe, 1997), while in others no impact of TIEs on comprehension was found (Cheng & Good, 2009; Jacobs et al., 1994, Jung, 2016; LaBrozzi, 2016). Several other studies confirm that TIEs are not necessarily beneficial for vocabulary acquisition and reading comprehension, and highlight the complicated relationship between the two. Rott (2007a) assessed and compared the combined/cumulative effect of word frequency, and semantic and visual enhancements on text comprehension and learning and long-term retention of vocabulary in three word intervention tasks. In the three treatment conditions, target words were (a) glossed four times, (b) glossed, L1 retrieved (through translation) and twice bolded; (c) glossed and three times bolded. After reading, the 38 participants were asked to retell what they had read, and then completed vocabulary post-tests measuring receptive and productive vocabulary knowledge. The study confirmed that,

depending on enhancement type, the processing of TIEs has a long-term effect on word learning. Semantic enhancements (glosses) were advantageous for developing productive vocabulary, whereas visual enhancements (bolding) had no clear effect. These studies suggest that readers, without instructions to do otherwise, do not direct their attention to lexical form but focus on comprehension, and that retention would require isolating the word form from context and processing meaning (Ellis, 1994; Rott, 2007; Hulstijn, 2001). This seems more likely to occur with glossing than with bolding. Rott's findings support the claims of the ILH that target words that are processed more elaborately are remembered better (Hulstijn & Laufer, 2001). This is in contrast to the findings from LaBrozzi (2016). He examined how L2 form recognition and reading comprehension are affected by different types of textual enhancement. Results of the study showed that typographic enhancement, especially increasing of font size, was an effective way of facilitating L2 form recognition. Learners considered this type of enhancement as "deliberate" and "significant" and therefore paid attention to this change in the text (p. 86). Comprehension was not affected by the enhancements, regardless of enhancement type.

Task Elaboration

The findings discussed above contradict to some extent the outcomes of Watanabe's (1997) study, which highlights task elaboration as the third of the initially mentioned factors of importance in this field of research. To explore the effects of increased processing and glossing, Watanabe investigated how different cue types would affect input processing, initial learning, and retention of target vocabulary. The 131 participants worked with ten types of TIE, including multiple-choice glosses, and restatements of difficult words as appositives within the text immediately following the target words. The participants also completed a translation task for more elaborate target word processing. Watanabe found that the glossed modification types lead to significant effects on immediate and delayed vocabulary learning, but not the appositives or the additional translation task. The higher elaboration did not lead to better results. It may be that the nature of the additional task (translation), while requiring an isolation of the word form, was so obtrusive, that it did not lead to a re-contextualisation of the processed word and a form-meaning connection could not be made. Otherwise, however, Watanabe's findings highlight the positive effect of glossing on vocabulary acquisition.

Rott (2007) found that reading conditions in which target words were either glossed four times or glossed and L1 retrieved and bolded twice lead to best results, which suggests that combined TIEs and frequency may lead to higher scores. However, her findings suggest that these different TIE types have differing effects: While glosses ensure correct word encoding in the mental lexicon, increased frequency seems to foster robustness (effects of frequency

are discussed in more detail below). Particularly interesting is Rott's finding that some vocabulary learning took place at the expense of text comprehension. As mentioned previously, findings from TIE studies in connection with text comprehension present a mixed picture. Despite evidence that TIEs promote the noticing of vocabulary without impeding comprehension (Izumi, 2002), the overall conclusion appears to be that TIE are useful, but not necessarily when used for aiding both vocabulary acquisition and reading comprehension simultaneously. Enhancements are more likely to induce learner noticing of the target form when sequential to comprehension than when concurrent with comprehension (Han et al., 2008). Both operations at the same time are difficult to master. Language processing research showed that processing both vocabulary acquisition and reading comprehension simultaneously overburdens learners' attentional capacity (see section 2.3.1). This is highlighted by research (Cho, 2010; Gettys et al., 2001) showing that reading for meaning and decoding unknown L2 words are two highly verbal tasks, "which are […] likely to interfere with each other and impede the process on the whole (Gettys 2001:93). The grammar-focus study by Winke (2013), however, questions that working with enhancements lead to a trade-off between comprehension and acquisition. Winke did not find evidence of enhancements detracting from comprehension, but by tracking learners' eye-movements, she found that enhancements had a significant positive effect on learners' noticing of the targeted grammar forms because of rereading and longer gaze durations.

Obtrusiveness

The concept of 'obtrusiveness' refers to the fact that glosses direct readers' attention to target items, lead to a de-contextualisation of this word, and may thereby interrupt the natural reading flow (Rott, 2007). Different levels of obtrusiveness refer to the degree to which TIEs lead to such interruptions. Obtrusiveness is based on the notion that certain reading behaviours display a 'switching cost', i.e. that handling limited cognitive resources, for instance when readers move between text and margin, potentially increases inefficiency and decreases productivity in task completion (Rubinstein et al., 2001). There is evidence that learners perceive some enhancement forms as obtrusive. Participants in Bell and LeBlanc's (2000) study said that the appearance of the page containing the gloss information and the necessity of returning to the page containing the text were distracting. Rott's (2007) study, however, provides evidence that levels of obtrusiveness and the height of switching cost also play a positive role for vocabulary acquisition. She found that more obtrusive word interventions, which stimulated the processing of the meaning of the word, for instance additional glossing or word retrieval, had a significant positive effect on the quality of word encoding and showed clear advantages for the development of productive word knowledge. LaBrozzi (2016) found

that the more obtrusive types of enhancement (increased font size) led to significantly better results than the more subtle forms (change in font). He assumed that these were more likely to go unnoticed by an L2 learner reading for meaning. However, in Rott (2007), tasks asking participants to interrupt reading for recalling and note taking led to inferior comprehension results. It might be that such tasks draw on attentional resources on the expense of comprehension. Alternatively, the learners might have perceived the task more as a word learning exercise than a reading task. Whatever the reasons might be, the inherent contrast between comprehension and vocabulary learning as two distinct skills must be remembered. Rott (2007) pointed out that the cognitive mechanisms involved in reading comprehension and lexical acquisition are not only different, but "they might even be in conflict!" (p. 166). On the other hand, Han et al. (2008) claimed that input enhancement is "premised on comprehension, its underlying thrust being to prompt occasional metalinguistic attention for the ultimate benefit of a balanced development in comprehension and acquisition" (p. 603; see also Doughty, 2004; Kim, 2006; Wong, 2005). The research results described above, however, do not show clearly how and whether this is possible. Accordingly, while results are not clear, caution is needed when using TIE for vocabulary acquisition and reading comprehension simultaneously.

These findings reveal that understanding the processes involved in the use of TIEs for vocabulary acquisition and reading comprehension is a challenge for research. In a small-scale study, O'Donnell (2012) explored how L2 learners' interaction with L1 and L2 glosses related to text comprehension. The analysis of her think-aloud-protocol data confirmed that gloss use and comprehension are related, but that it is not clear how. It was, for example, difficult to determine whether gloss use influenced comprehension or whether it worked the other way around, i.e. that better comprehension endorsed successful gloss use. Regarding TIE use, whether L1 or L2 glosses were used did not seem to matter, as both forms were handled with a certain degree of confusion. O'Donnell recommended L1 glosses for early learning stages and L2 glosses only when learners are sufficiently proficient.

O'Donnell's (2012), Cheng and Good's (2009), and Rott's (2007) studies show that still little is known about the vocabulary learning processes involved when TIEs are used for comprehension. Therefore, in their 2008 article on critical issues and possibilities of TIE, Han et al. claim that when investigating TIE efficacy, in addition to noticing, comprehension needs to be measured. Given the constraints on attention distribution discussed earlier, this seems a sensible claim: through TIE, learner attention may be directed to the target forms potentially at the expense of global or local text comprehension (e.g. Shook, 1999; Rott, 2007). This is not desirable and it seems appropriate to follow Han et al.'s claim. However, while I included reading comprehension in my study, I did not measure this skill systematically enough to carry out a statistical analysis.

Simple versus Combined or Compound Enhancements

Another TIE study relevant in the context of research investigating vocabulary learning from reading enhanced texts is Kim (2006). He focussed on incidental vocabulary acquisition from reading texts in which target words were either typographically enhanced (bold printing), lexically elaborated (either implicitly or explicitly), or both. Reading texts that were only typographically enhanced did not significantly change scores in form or meaning recognition tests, and neither did enhancements in the form of explicit/implicit lexical elaborations. However, for texts where lexical elaboration and typographical enhancements were combined, significant effects were recorded for meaning recognition. It seems that combined forms of enhancements are to be recommended.

The potential superiority of combined forms of enhancements was confirmed by several other studies, among them Tabatabei and Shams (2011). They tested 60 high-school students to investigate effects of different types of multimedia glosses (picture, L2 definition, or a combination) on online reading comprehension and vocabulary acquisition. They found that students working with any of the glossed texts significantly outperformed the control group, which worked with unenhanced material, in both comprehension and vocabulary recall. Double glossed texts (picture + definition) led to better results in both aspects than the texts containing enhancements of only one type. Similar findings were made by Jourdenais (1998, 2001), Kost et al., (1999), Min (2008), Yoshii (2006) and Zandieh and Jafarigohar (2012). All of these studies confirmed that a combination of different enhancement forms led to superior results.

Several studies investigated ‘compound enhancements’, i.e. combining TIEs with other salience-increasing forms of enhancements, for instance additional exercises, pictures, or explicit teaching. Kost et al. (1999) investigated effects of pictorial and text glosses. They found that the superiority of compound enhancement stems from a ‘dual coding of the input’, i.e. that a combination of the two instruments allows readers to “store new information in two different manners, i.e., in verbal and nonverbal storage systems”, which increases the reader’s number of retrieval options (p. 95; see also Lee, 2007). Grammar-focussed TIE studies, for instance Izumi (2002) and Williams (1999), also showed that input enhancement works best in *combination* with other enhancement types. It appears that visual highlighting alone is often not enough to prompt learners to go beyond simple forms detection, and that additional (instructional) assistance is needed. Laufer (2001) found similar success for post-reading tasks in combination with input enhancement. In her study, the readers who engaged in pedagogical activities after having read the text and who were provided with the word meaning outperformed readers who received an unenhanced text. Her findings highlight the possible

positive effect of pedagogical pre- and post-reading tasks on word learning and encourage a holistic approach to input enhancement.

Prior Topic Knowledge

Other grammar-focussed studies found that the effectiveness of simple or compound enhancements depends on the learners' knowledge of the target forms and affects the level of knowledge that can be acquired. Simple enhancements are more effective for learners with some prior knowledge than for learners without (e.g. Alanen, 1995). Prior knowledge is the sixth of the eleven factors that contribute profoundly to the effects of TIEs.

Related research investigated the effect of background knowledge and found strong effects of topic familiarity for instance on lexical inferencing success (Pulido, 2003; 2007), and on how learners cope with quasi-dual-task situations, in which they must attend to form and meaning simultaneously (Lee, 2007). As for TIEs, simple enhancements may induce noticing but not understanding in learners with little prior knowledge (Shook, 1994). However, they may incite comprehension as well as noticing in learners with some prior knowledge (Lee, 2007). Compound enhancements are superior to simple enhancement in inducing noticing, and further processing of the target form in both types of learners (Williams, 1999). This learner familiarity with the target forms has been found to be a decisive factor affecting the potential success or failure of TIE in several studies. Waring and Takaki (2003), in their study of reading of a graded reader, concluded that reading a graded reader was more instrumental to enhancing and enriching already known vocabulary than to facilitating new vocabulary acquisition and retention. Izumi's (2002) study underscored that, when meaning is clarified before a focus on form, input enhancement is likely to stimulate even growth in comprehension and acquisition. These findings are important for using TIE effectively in teaching.

Language Used For Glossing

Another approach to investigating the effectiveness of different TIE forms was to test the effect of the language used for glossing (e.g. L1 or L2). Bell and LeBlanc (2000) examined whether L1 or L2 glosses were better suited for text comprehension. Working with computerized texts, they investigated learners' attitudes towards L1 versus L2 glossing. Vocabulary uptake was not measured. They found an overwhelming preference for L1 glosses and their computer tracking data revealed that L1 glosses were consulted far more often than L2 glosses. The language of the glosses did, however, not significantly impact on reading comprehension. This is in line with Jacob's (1994), Ko's (2012), and Yoshii's (2006) findings. Laufer and Shmueli (1997) discovered that L1 glossed words were always retained better than words glossed in

the L2. They also compared the effects of different modes of presentation of the target words (isolated or within context, within a text passage, or in a list). The results showed that in all forms of presentation, working with the L1 led to superior scores. The researchers concluded that providing L1 one-word-glosses enables learners to pay full attention to new L2 words since the L1 words are already familiar. Similarly, Yoshii (2006), exploring how effectively L1 and L2 text and picture glosses support incidental vocabulary acquisition in a multimedia environment, found that both L1 and L2 glosses fostered incidental learning effectively, but that L1 glosses may have a more long-lasting effect. Similarly, in Fahimipour and Hashemian's study (2013), L1 glosses led to results superior to those from working with L2 glosses.

Internally Versus Externally Imposed Enhancements

As mentioned before, it is assumed that the efficacy of TIEs is primarily controlled by the learner and not by the researcher (Han et al., 2008). Sharwood Smith (1991) hypothesised that enhancements chosen by readers, i.e. internally initiated enhancements, are more effective, than externally imposed forms chosen by the provider of the reading materials, for instance a teacher. This is because with the former type, learners force themselves to engage actively with the target forms based on their own needs and purposes. The two grammar-focussed studies of Leow et al. (2003) and Shook (1999) confirmed this and thus lent empirical support to Sharwood Smith's (1991) notion of 'learner-created salience'. Whether and how learners choose items for enhancement depends on the items' communicative value, salience and the learners' need to gain knowledge about them, as Leow et al. (2003), found. They compared different types of TIE and concluded that input flood – where learners decide which items to focus on has a greater attention-getting capacity than externally chosen enhancements. According to them, the importance of learner-created salience can supersede that of externally chosen items. Han et al., (2008) assume that the problem with merely externally enhanced input may be that the learners either don't notice it, or that they notice it only partially, contingent on whether or not they are ready for it or how much overlap there is between externally and internally generated salience. As Sharwood Smith (1991) puts it: "A mismatch may arise between the intentions lying behind teacher or textbook generated enhancement of the input and the actual effect it comes to have on the learner system." (p. 130). While the investigation of externally imposed enhancements is at the heart of my study, the role of internal enhancements is investigated in my interviews.

The finding that the language items that learners themselves find relevant, are noticed and processed more readily then externally enhanced items, confirms the tenets of the ILH (section 2.3.5). Several TIE-related studies point in this direction. Han et al. (2008) pointed out, that,

for the target forms to be meaningful[15] is vital, in the words of the ILH, that there is a *need* for the target item, and added that enhancing a non-meaningful form does not lead to better intake. Hulstijn et al. (1996) examined whether combining a heightened frequency of occurrence of target words and providing word meaning in glosses or through dictionary use would have an effect on the incidental learning of L2 French words. They found that the gloss groups recorded better attainment than the dictionary- and the control group. Relevant when considering the factor *need* is Hulstijn et al.'s (1996) finding that while the glosses had a greater effect, the participants used their dictionaries only when absolutely necessary for comprehension. This may suggest that glosses are used beyond what may be needed because they provide such easy access to information. More support for the ILH was provided by Min (2008), who concluded that elaboration of input enhancement makes a difference. In Min's study, the 'reading group' recalled significantly fewer words than the 'sentence writing' and 'sentence completion' groups, i.e. the groups working with higher elaboration-tasks.

Frequency as Enhancement

Another central aspect when investigating different TIE forms is the role of frequency as a form of input enhancement. It seems logical that the repeated occurrence of a new word would be a predictor for word learning. Several studies have shown positive effects of a heightened number of encounters with words in reading tasks (Eckerth & Tavakoli, 2012; Révész & Brunfaut, 2013; Rott, 1999, 2007a; Waring & Takaki, 2003; Zahar et al., 2001). Rott (2007), for instance, found that enhancements linked to heightened frequency of exposure could lead to superior text comprehension (see also Davis, 1989; Jacob, 1994; Watanabe, 1997). To be retained through incidental learning, an unknown word needs to be encountered many times (15 times according to Waring & Takaki, 2003), and even then the acquisition of the word cannot be guaranteed and long term retention is weak (Eckerth & Tavakoli, 2012). As frequency is such an important aspect, I considered to include this factor in my study. However, I decided not to use frequency, as the texts in my study are too short to repeat target words often enough for frequency to have a measurable effect.

TIE Non-success

As seen above, some TIE-forms work best for specific purposes and in certain circumstances. However, TIE does not always work. Several studies explore the reasons why working with enhancements might fail to lead to the desired effects (e.g. vocabulary learning or

[15] Meaningful means 'of relevance for task completion', e.g. reading comprehension.

comprehension). This is the tenth aspect I mentioned in the list of important factors for understanding TIEs. Han et al. (2008) pointed out that the assumption that TIE draws learners' attention to the target form, which, in turn leads to further processing of it, may be too generic. It is problematic to expect instant learning to arise from a single (typically, short) treatment session of TIE.

Izumi (2002) investigated the effects of TIE and output exercises on noticing and learning of problematic form features. Despite a positive impact of TIE on the noticing of the target items, participants who had received enhanced materials failed to show measurable learning gains. Izumi reckons that input enhancement may not be enough to induce noticing of the interlanguage–target language mismatches. Furthermore, he assumes that enhancements, which are solely concerned with drawing learner attention to form, do not necessarily encourage further cognitive processing necessary for acquisition. Accordingly, TIE may have caused mere recirculation or rehearsal at the same, relatively shallow, processing level, which led the learners to experience only a short-term retention of the attended forms. This shows that input enhancement does not necessarily lead to deeper processing and therefore is not necessarily boosting acquisition. Discussing similar results, Rott (2007) points out that encountering an unfamiliar word in a text or comprehending it in context might not lead to an initial assignment of meaning to word form.

Certain conditions are needed if TIEs are to achieve raising salience, aiding noticing and ultimately acquisition. Not all of these have been researched, but some have become apparent. First, input enhancement has to be integrated into the materials in a logical way. For example, bolding individual words in a text without indicating the purpose might cause readers to attend to what word aspect they consider most important (e.g. spelling) or to not attend to any particular aspect at all (Rott, 2007). This was apparent in White's (1998) study where participants were not certain about the purpose of the enhanced forms and did not react to them in the desired way. Secondly, some studies focussing on grammar showed that the enhanced forms need to be of communicative value[16]. Studies whose target forms have a higher communicative value (Izumi, 2002; Lee, 2007; Song, 2007) revealed positive effects of TIE, while those with low communicative value showed no effects (White, 1998; Wong, 2003).

Accordingly, it is possible that the positive results of TIE in some studies are due partly to the particular selection of target forms. Williams and Evans (1998) found that not all forms benefit equally from TIE. This is likely to be true for vocabulary as well. For example, Laufer (1989, 1997, 2013; see also Laufer and Shmueli, 1997) discussed how intra-lexical factors

[16] According to VanPatten (1985), communicative value refers to the meaning that a form contributes to overall sentence meaning and is based on semantic value.

may affect word learning, and Webb (2008) found quality of context to influence meaning acquisition.

Learning Collocations

Recently, several TIE-related studies investigated how enhancements affect the learning of collocations and similar multiword units and found beneficial effects (Boers et al., 2016, Choi, 2016, Peters, 2012). This seems to be an obvious point of interest, as, according to Boers et al. (2016:2), knowledge of collocations is "one of the hallmarks of native speaker competence", while acquisition is slower than for single words. For mulitword units in particular, one encounter is insufficient, as the word string might not be noticed as a linked unit. Using enhancements can be especially helpful with this.

Peters (2012) examined the effect of typographic enhancements and of instructing learners to pay attention to formulaic sequences and single words on vocabulary recall. Typographic salience affected participants' recall scores positively, whereas the instructional method did not. The effect of typographic salience seemed to be particularly beneficial for learning the formulaic sequence as it facilitated noticing. Boers et al. (2016) compared the effect of different typographic enhancements on the intake of multiword units from reading. Their findings showed that the enhanced multiword units were more likely to be recognized than unenhanced ones. It seemed, however, that the extra attention paid to the enhanced items "usurped attention at the cost of others" (Boers et al., 2016:12). Similar results were found by Choi (2016), who conducted an eye movement study investigating the processing and learning of enhanced English collocations and testing text recall. The study's outcomes indicated that better performance in the post-test was due to the target collocations being processed longer. However, they also suggest that recall of unenhanced text suffered.

Long-Term Effects

Finally, while finding the most effective TIE-forms is important, so is to investigate whether these effects facilitate long-term retention. Research did not produce clear findings. Rott (2007) found that the processing of glosses had a long-term effect on word learning, albeit not on reading comprehension. Zandieh and Jafarigohar (2012) compared reading comprehension and vocabulary acquisition from intentional versus incidental reading. In delayed tests, the incidental group reported a slightly higher retention rate and an overall more stable vocabulary gain. Ko (2012), comparing effects of L1 versus L2 glosses, administered a four-week delayed test. The results showed significant differences between the glossed and the un-glossed conditions, but not between the two types of glosses. Cheng and Good (2009)

found a decrease in vocabulary recall scores in a one-week delayed test and then a slight increase in the two-week delayed test. The repeated review of glossed words might have led to this increase. The overall tendency that long-term effects on vocabulary acquisition can be measured across different types of glosses and reading types was also confirmed in several of the studies reviewed by Abraham (2008; see also Huang, 2003; Watanabe, 1997).

In conclusion, it seems that directing L2 readers' attention to new words by providing TIEs can improve the likelihood of establishing and strengthening lexical form-meaning-connections. However, the research discussed in this chapter shows that working with TIEs does not necessarily lead to these effects. It remains unclear which types of TIEs work most effectively and for which purposes, as the learning situations, as well as the methodologies reported in the studies differ in several aspects. Still, despite the disparity in the above-mentioned studies, there are common aspects that seem to show the effectiveness of using TIEs for vocabulary acquisition.

One of these decisive aspects is the combination of various forms of enhancements and linking enhancements with related activities. Other aspects to be considered are learner proficiency and familiarity with the target forms. More advanced learners work more efficiently with certain types of TIE and previous knowledge of the target forms supports TIE success. Moreover, if text difficulty goes beyond the learners' proficiency level, TIE may be used erroneously or ignored. Linked to proficiency, another decisive finding was that L1 enhancements, as opposed to L2 enhancements, tend to lead to higher scores in both reading comprehension and vocabulary acquisition.

Additionally, the form of TIE is a crucial aspect. In all considered studies, semantic enhancements (e.g. glosses) lead to results superior to visual enhancements (e.g. bold printing). Whether, for example, simple glosses, multiple-choice glosses or sentence-length glosses are most effective depends on various factors. It is assumed that the height of 'switching cost' or the 'level of obtrusiveness' comes into play here. Beyond this, the type of items chosen for enhancement is vital. It was found that items that learners feel the genuine need to engage with, paired with the forms which encourage deep processing, will most likely be noticed and processed. These are items where the internal drive of the learner and the concerns of the teacher (external) overlap.

Ultimately, it is the learners who decide how to approach a language learning tasks. What appears to be relevant to them and what does not, therefore, determines the outcome of their learning success. Therefore, over the years TIE-research has become increasingly interested in learner attitudes. Studies that explore how learners feel about enhanced texts are presented in the following section.

2.6.2. Research On Readers' Attitudes Towards Input Enhancements

One of the first TIE studies was Holley (1970, quoted in Holley & King, 1971). In this study, readers of enhanced materials filled in a questionnaire regarding their attitude toward working with TIEs. The results show clearly positive attitudes towards TIEs in general and glosses in particular. Many of the participants believed that glosses helped and accelerated vocabulary acquisition. Similar positive attitudes were reported in several studies since (Bell & LeBlanc, 2000; Cheng & Good, 2009; Ko, 2012). Recent research gave further insight into learner attitudes. Readers favour glosses that are concise and easily accessible and, if given options, they prefer glosses which provide definitions of unknown words, rather than grammatical explanations (O'Donnell, 2012). O'Donnell also found that if presented with options, L2 readers tend to select L1 glosses over L2 glosses. Participants stated, however, that they would prefer to use comprehensible L2 glosses. Similarly, in Ko's (2012) study, learners preferred L2 glosses over L1 glosses. As mentioned above, which types of glosses are used and preferred is tightly linked to learners' proficiency level. Learners would prefer L2 glosses if they felt confident to understand them sufficiently well (see also Jacobs et al., 1994; Taylor, 2006). Ko's survey data further revealed that 92% of the participants preferred glossed texts over un-glossed texts, mostly because this facilitated reading as such, aided guessing and vocabulary acquisition during reading (see also Hulstijn, 1992; Jacobs et al., 1994). The small percentage of participants (8%) who preferred reading unenhanced materials mentioned that "glosses can make learners lazy" (Ko, 2012:74). They felt that TIE negatively influenced their habit of consulting dictionaries. Additionally, Ko revealed that L2 glosses were the most preferred enhancement form and, while most readers concentrated on comprehension rather than vocabulary acquisition, they still consulted 70% of the glosses. This behaviour and the question of whether glossing really makes learners 'lazy' are discussed in the interpretation of my interview data.

Another relevant aspect was revealed by Jacobs et al. (1994). Participants indicated a near unanimous preference for marginal glosses over those placed in locations more distant from the text. Learners prefer convenient, easy to access enhancement types with a low level of obtrusiveness, so that they can focus on comprehension, and understanding/acquiring vocabulary can happen incidentally. This was confirmed in Gettys et al's (2001) study, which compared sentence-level and dictionary-form glosses. Although vocabulary learning was higher when students read the text with dictionary forms of the unknown words, most readers preferred sentence-level glosses. Participants explained that sentence-level gloss definitions were easier to access, more direct, and facilitated more thorough understanding.

In addition to learner attitudes, TIE research investigated aspects such as the density and location of enhanced items, or whether presenting reading materials on a computer screen or

on paper plays is influential. However, so far these aspects have not revealed substantial effects and are therefore outlined only briefly below.

2.6.3. Other Factors Researched In This Field

While placing glosses in the margin is the most common form of enhancement, footnotes or word lists at the bottom of a page or the end of chapters/texts have been used as well (e.g. Holley & King, 1971). Research studied learners' responses to those different gloss locations, but, as far as I am aware, their differential effects on vocabulary acquisition have not been measured. However, the ideal ratio of unknown enhanced words in a text, i.e. enhancement density, was investigated. However, no definitive answers were found, as previous studies have not been consistent in selecting and limiting the number of target glosses. Studies addressing these issues do so by collecting qualitative data or by pointing out that location and density are issues that ought to be considered (Ko, 2012; Taylor, 2002).

Holley and King (1971) were interested in whether enhancing reading materials with glosses in different locations (margin, bottom of page, in attached word list) had an effect on vocabulary acquisition and comprehension when reading texts with different densities of new words. Vocabulary acquisition and comprehension of 110 students of German were measured with multiple-choice tests. The findings showed no significant effect of locations like glosses in the margin or word lists, and no significant effect of target-word density.

Gloss density and location were also considered by Bell and LeBlanc (2000). A questionnaire elicited participants' preference for higher or lower gloss density. 80.5% reported a desire for vocabulary help with additional words. As mentioned above, Jacobs et al. (1994) inquired about gloss-location preferences and found an overwhelming preference for glosses placed in the margin. They also considered gloss density and found glossing to be effective for vocabulary acquisition; this was likely to be the result of high gloss density chosen in their study. A much higher than average number of vocabulary items had been glossed (53 out of 483 words). Jacobs et al. assumed that researchers in other studies had underestimated the amount of help students needed. The exact 'average' number of glossed items in other studies, however, was not reported. Other researchers suggested to investigate which amount of glosses was best for which proficiency level (Ko, 2012).

To date there is little research considering the effects of location and density of glosses. It is likely, however, that soon both aspects will be investigated in more TIE studies, as recently there has been a clear tendency to focus on the role of reading enhanced texts online. With the various possibilities of manipulating texts on a computer, it is likely that glossing in general and location and density will receive particular attention in research.

The increasing use of technology has led researchers to consider TIE in the context of reading on screen versus on paper, but only few studies explored whether the type of presentation actually made a difference regarding comprehension and vocabulary acquisition (Bowles, 2004; Gascoigne, 2006). Several studies tried to find out whether computerized TIEs offer better learning opportunities (Abraham, 2008; Lomicka, 1998; Rashtchi & Aghili, 2014). Roby (1999), for example, discusses the technological possibilities with respect to both location and density. When I planned this study it was, unfortunately, not possible to include online reading materials, so that my data collection was based on texts read on paper.

My review of TIE literature above should have made it obvious that there are several TIE-related issues that have not yet been sufficiently addressed by research. The most pressing ones and the resulting research questions are presented below.

2.6.4. The Contribution Of My Study

My review of TIE research points to a multitude of issues that still need to be investigated to better understand word learning from reading texts with enhanced target words, and how this acquisition process can be made more effective for the benefit of the learners. This study attempts to fill some of the knowledge gaps. Below, I first mention the issues that need further investigation. Then I explain which of these I address in my study.

- There already are many studies that try to identify the most effective TIE-types, but findings are confusing. Therefore, adding knowledge to this body of research is desirable. As the mixed picture, among other things, stems from the wide variety of methodologies, TIE types, learning contexts, text lengths, studies that replicate existing ones ought to be encouraged. This would allow more coherent comparisons between findings. Therefore, I use TIE-types that are not only common in teaching materials, but have been used in previous studies. Furthermore, I work with text lengths comparable to those used before.

- There is still a discussion about the benefits of simple versus combined enhancements. By including both types, my study will help refining the findings.

- Following several previous studies, my vocabulary acquisition assessment tools measure active and passive forms of both receptive and productive word knowledge. Thus, I hope to make my findings comparable to existing research and gain new insights into how TIEs influence different facets of vocabulary knowledge.

- Most previous studies were conducted in academic learning environments. However, it is assumed that the largest number of instructed language students learn in a school context. Situating my research in authentic language learning classrooms in high-schools instead of an academic context, and testing younger and less proficient language learners fills a knowledge gap and was therefore put into action in my research project. This removes a potential academic bias in the findings and makes my contribution highly relevant for a wide range of L2 learners.

- There is little research which investigated the concepts of 'obtrusiveness' and 'switching cost' and their effect on word learning even though previous research has found them to be potentially influential (Lee, 2007; Wong, 2003). Several researchers have highlighted this gap (e.g. Rott, 2007; Watanabe, 1997). One goal of my study is to contribute to research and theory building in this area.

- Additional studies investigating not only TIE's immediate effect on vocabulary learning, but also its intermediate and long-term effects are necessary. For that reason, I included two delayed post-tests in my research design.

- Exploring how learners interact with TIE is another issue that has been considered by only a few studies. However, the aforementioned concern that the effectiveness of input enhancement may not be as much controlled by the researcher as by the learner suggests that more research is needed. My interview data is hoped to add to existing knowledge.

- It is still unclear what the relation between 'incidental' and 'intentional' learning is like and whether the two are two distinct categories or overlapping. There is so far only little empirical evidence showing what learners focus on in incidental reading and in particular, how they perceive and approach reading tasks containing TIE. New knowledge about this may help linking the theoretical constructs incidental/intentional to the reality of learning situations.

- There has been little research investigating the specific amount of enhanced words most suitable for different proficiency levels, but several researchers have suggested this as a point to consider. It would also be useful to examine the effects of TIE in a L2 reading theoretical framework, as there is little research investigating how TIE interacted with reading processes. Unfortunately, it was impossible to do this in my study.

The following research questions emerged from my literature review:

(1) *What are the immediate and long-term effects of different types of textual input enhancement on incidental and intentional learning of EFL-vocabulary from reading tasks?*
(2) *Does textual input enhancement have a differential effect on different types of word knowledge?*
(3) *Which task-related and learner-internal factors emerge as relevant for vocabulary learning from reading enhanced texts and how do these affect how learners use input enhancements in reading tasks?*

The following chapter describes the methodology that was used to investigate these questions.

3. METHODOLOGY CHAPTER

3.1. Introduction

This chapter explains the methodology used to investigate the research questions and the chosen research design. In section 3.2, I describe the qualitative and quantitative research approach and discuss the possibilities and limitations of these two paradigms. Then, the concepts of validity and reliability are explained. The chapter continues with a discussion of the contribution that the triangulation of methods can make to research and ends with a presentation of the overall methodological approach chosen for this study.

Section 3.3 describes the research set-up, the data collection, and the participants. Ethical issues are explained in section 3.4. In 3.5, I explain the choice of texts, target words and the different types of enhancements. Section 3.6 explains the function and structure of the reading tasks, the instructions, as well as the reading comprehension questions and pre- and post-reading activities.

I conducted two pre-tests to find the most suitable target words and test format. Section 3.7 explains the objectives and procedure of these tests and comments on their limitations and what measures were taken to confine them. The pre-test results led to several adaptations and were the basis for the tests that were finally chosen to measure vocabulary learning. Form and purpose of the delayed vocabulary post-test are explained in the last part of this section. Section 3.8 clarifies the aims and form of the interviews. In the pilot study, the chosen texts, target words, test format and procedure as well as the interviews were tested with a group of learners similar to the sample group. Issues concerning this pilot study are described in Section 3.9. The following section reports how the main data collection was conducted. Finally, Section 3.11 indicates how the data collected in the main study were analysed.

3.2. The Methodological Approach

How data are collected and analysed may reflect an ideological orientation and depends on the topic under investigation. Most social science studies can be more or less suitably categorized as following either a quantitative or a qualitative approach to empirical research (Angouri, 2010). Even though it has become more and more common in SLA-research to mix the two approaches (e.g. O'Donnell, 2012; Rebuschat et al., 2015), the 'quantitative' and 'qualitative' distinction is still relevant. The nature of these two paradigms is therefore outlined in the following section, and the implications of triangulating different approaches are presented thereafter. In addition, the two concepts of validity and reliability need to be addressed, as they are highly relevant in all these domains. Lastly, it is explained which approach was chosen for my study.

3.2.1. The Quantitative And The Qualitative Research Paradigm

The most significant interest of quantitative research is to measure common features of (large) groups of people, therefore standardized procedures, which are stable across investigators and subjects, are needed (Bryman, 2012). Within this approach everything is expressed by numbers, which is an attempt to minimize the influence of researcher bias or prejudice and to make findings generalizable, for example through in-built quality checks, such as statistical significance. To guarantee that an investigator's partiality does not influence the outcome, quantitative researchers define the categories under investigation well in advance and assign a logical scale of values to them, which can be communicated by numbers. The reliance on numbers opens up possibilities, but also sets limitations. Numbers could mean anything without a contextual backing: in the case of this study, vocabulary tests show gained knowledge of the assessed target words, but reveal only little about the nature of the learning process. In quantitative research responses are averaged out, so that it is impossible to do justice to subjective variety. However, "similar scores can result from quite different underlying processes, and quantitative methods are generally not very sensitive in uncovering the reasons for particular observations or the dynamics underlying the examined situation or phenomenon. That is, the general exploratory capacity of quantitative research is rather limited" (Dörnyei, 2007:35).

For a long time, research in applied linguistics was dominated by the quantitative approach, which was regarded as the more 'scientific' paradigm (Holliday, 2010; Lazaraton, 2000; 2005). In reaction to this attitude, which does not account for individual cases, in the last decades more and more scholars in language related studies started to use qualitative methods, such

as the ethnographic narrative, interviews and observations to find answers to their research questions. Qualitative research is different from quantitative research in two important ways: it is not numerical but verbal; it is usually not determined a priori but open-ended and flexible as long as possible to be able to account for the subtle nuances of meaning uncovered during the investigation (Holliday, 2010, Dörnyei, 2007). Therefore, qualitative research is often seen as 'emergent'; i.e. the study is kept open and fluid so that it can respond flexibly to new details that may emerge during the process of investigation. Unlike quantitative research with its predetermined categories, qualitative research "highlights the importance of the complexity of individuals and individual situations" (Creswell, 2014:4). As a result, qualitative data sampling is more complicated. The aim is to "search for the richest possible data" (Holliday, 2010:101), i.e. the researcher must "find individuals who can provide rich and varied insights into the phenomenon under investigation so as to maximize what we can learn. This goal is best achieved by means of some sort of 'purposeful' or 'purposive' sampling" (Dörnyei, 2007:126).

Qualitative research can reveal subtle elements of a matter that, due to the exclusive reliance on numbers, are lost in quantitative research. However, qualitative methodology has its own intrinsic shortcomings. First, the underlying 'tabula rasa' idea that any researcher goes into a project with a totally free and unbiased mind is naïve. Given that qualitative research is fundamentally interpretative, this may be problematic. However, today the field is "moving increasingly towards a postmodern acknowledgement of the inevitability of qualitative research being subjective" (Holliday, 2010:98). To disclose researcher bias, Creswell (2014) suggests that researchers should not only reflect on their own biases, values, and assumptions, but also actively write them into their research (p. 18).

Due to the time consuming nature of qualitative research and because this type of research is inherently concerned with characteristics of the individual, sample sizes are usually small. This makes the generalizability of findings from qualitative research another problematic, yet important issue to consider. The specific conditions found in one particular situation may not apply to others.

The choice of a methodological approach depends on the nature of the research questions. Generally, it seems that qualitative research is advantageous for exploring new areas, for making sense of complexity, and to investigate social phenomena as they occur naturally. Also, qualitative research is good for exploring language, because "every aspect of language acquisition and use is determined or significantly shaped by social, cultural, and situational factors" (Dörnyei, 2007:36). Quantitative research methods on the other hand are better suited for examining trends and for clearly distinguishable and measurable variables.

The qualitative-quantitative distinction is sometimes too crude and seeing research designs as purely either or is too simplistic (Nunan, 2005). Rather, it has been suggested, the

"relationship is best thought of as a continuum of research types" (Mackey & Gass, 2005:2). The two approaches can overlap; for instance by defining categories through which certain aspects of qualitative data can be quantified qualitative research can be 'turned into' quantitative.

3.2.2. Validity And Reliability

Both qualitative and quantitative researchers need to demonstrate that their studies are trustworthy. Particularly within the quantitative paradigm, with its pre-determined categories, the construction and precision of the measurement instruments is essential. Within the qualitative paradigm, the researcher is the instrument, and precision depends on the ability and effort of the researcher. To ensure both types of precision two aspects, reliability and validity, are in focus.

Within the quantitative framework, reliability is concerned with instrument consistency. The question is if instruments measure what they are supposed to measure. This is significant to ensure replicability or repeatability of results (Golafshani, 2003). This aspect is closely related to the generalizability of research findings: If the results are reliable, they are consistent over time and an accurate representation of the total population. However, reliability relates not only to the instruments themselves ('internal consistency'), but also to the researcher ('rater reliability'). It is always important to make sure that the definition of an item/construct (e.g. 'vocabulary knowledge') is sufficiently specific to allow any researcher to identify them as such (Mackey & Gass, 2005). Similarly, 'interrater reliability', i.e. the measure of whether different raters judge the same set of data in the same way, and 'intrarater reliability', i.e. would the same researcher evaluate the data in the same way at different times, have to be ensured (Mackey & Gass, 2005).

Validity deals with the question of "whether the research truly measures that which it was intended to measure" (Joppe, 2000:1; cited in Golafshani, 2003:599). Validity applies to the *content* range of the investigated item, i.e. the testing instrument needs to be sensitive to the full range of the item that is investigated (e.g. vocabulary knowledge). *Construct validity*, which refers to the degree to which the research adequately captures the construct of interest, (Mackey & Gass, 2005) is important. Particularly in SLA, because many of the variables under investigation are not easily defined. My study will specifically try to address the aspect of *face validity*, the participants' perception of the research treatments and tests (attitude, how familiar are the students with this type of task, do they know what to do; will they take it seriously) by investigating these points in the interviews. Concerns about *external validity* (random sampling, representativeness) and *internal validity* (participant characteristics like proficiency, language background, and learning experience) in this study are discussed in 3.3.

Both reliability and validity are mainly quantitative concepts; in qualitative research the two concepts are considered differently. According to Creswell (2014) validity is “based on determining whether the findings are accurate from the standpoint of the researcher, the participant, or the readers of the account”, i.e. it is concerned with trustworthiness and credibility (p. 201).

Whether reliability plays a role in qualitative research at all, is a controversial issue. Some scholars (e.g. Glesne, 1992; Winter, 2000) think that replicability is less relevant than precision and transferability. Other qualitative scholars, however, claim that reliability depends on validity and the latter, therefore, is sufficient (Golafshani, 2003). To guarantee reliability in qualitative research, Creswell (2014) proposes to document meticulously all taken steps and to compile a detailed “study protocol”, so that others could follow the procedures easily (Creswell, 2014:203).

Trustworthiness or rigour need to be warranted. Triangulation and using interraters are good strategies to control bias, for improving validity, and reliability of evaluation of findings.

3.2.3. Mixed Method Research And Triangulation

Both research approaches have something to offer for answering research questions. Often, the same research question can be investigated by complementing different methods, so that new aspects of the same issue are uncovered. As Morse and Niehaus (2009) state, “the defining characteristic of mixed method research is that it involves a primary or core method combined with one or more strategies drawn from a second, different method for addressing the research question” (p. 14). Working with multiple methods offers several advantages. Different data types bring issues to light that could not have been found by using one sort of data alone. Triangulation is also a way of validating hypotheses by examining them through multiple methods. As all methods have biases and weaknesses, collecting both types of data “neutralizes the weaknesses of each form of data” (Creswell, 2014:15), at least to some extent. Mixing methods can also be a way of reaching multiple audiences – scholars who usually follow only one paradigm might be willing to consider studies, which integrate the one approach in their methodology without completely abandoning the other (Dörnyei, 2007).

When researching a phenomenon as complex as vocabulary acquisition and the circumstances linked to it, combining multiple types of data and measures is beneficial, as “any single measure of it will give a very minimal impression of the overall lexical knowledge constellation” (Schmitt, 2010:152). The question remains how the qualitative-quantitative combination can be achieved. How methods are applied in order to be truly meshed and whether one or the other method should be dominant, must be continuously determined by the “theoretical force” within the project (Morse & Niehaus, 2009:11).

3.2.4. Approach Chosen For My Study

My research questions propose the investigation of the following variables: vocabulary acquisition as the dependent variable, and textual input enhancement (TIE), task approach and other task-related and learner-internal factors as the independent variables. My study aims not only at finding out how these variables affect vocabulary acquisition, but also how they influence each other. This requires both quantitative and qualitative methods.

To measure the effects of different TIE-types on vocabulary learning I collected vocabulary post-test data, which were analysed statistically. To gain insights into the underlying learning processes, retrospective learner interviews were conducted. I hoped that relating the findings from the interviews to the results from quantitative data analysis could reveal the relationship between these variables. Details that the quantitative tests do not reveal can be explored in the interviews, so that a more holistic picture of vocabulary acquisition can be achieved.

According to Tashakkori and Creswell's (2007) categorization, the research design chosen for my study integrates the two big paradigms, as it combines two types of data collection procedures (vocabulary tests and retrospective interviews); two types of data (numerical and textual) and two types of data analysis (statistical and thematic). Specifically, I use a "convergent parallel mixed methods" design (Creswell, 2014), a mixed methods design in which "the researcher converges or merges quantitative and quantitative data to provide a comprehensive analysis of the research problem" (p. 15). This approach is useful when the quantitative or quantitative approach, each by itself, is inadequate to fully understand a research problem.

To assess the different facets of the acquired vocabulary knowledge, I followed previous TIE-related studies (e.g. Bell, 2000; Gascoigne, 2006; Laufer et al., 2004; Rott, 2007), in which quantitative data was collected and analysed statistically to measure and compare vocabulary gains resulting from various forms of TIE. This approach seemed suitable not only to make my findings comparable to previous research, but also because findings from the quantitative analysis would allow me to make generalizable statements for all groups of participating students. These results could be easily categorized (e.g. type of TIE used, active versus passive recognition, target word scored correctly/incorrectly) and compared within and across groups. Moreover, the statistical analysis enabled me to some extent to make statements about causal effects and investigate the correlation between the different variables. However, investigating the benefits or shortcomings of using TIEs for vocabulary acquisition, as was done in my study, goes beyond what is measurable with numbers only, for instance when enhancements are used in unexpected ways.

As my study was explorative, it was neither possible nor desirable to control for all factors that impact vocabulary learning from reading enhanced texts. Rather, the intention was to identify other influencing factors and to investigate their impact on the learning process. Therefore, by using interviews rather than for example questionnaires, I added to the knowledge gathered through quantitative research with a qualitative approach. Interviews are flexible and allowed me to investigate phenomena that are not directly observable, such as learners' self-reported perceptions of their use of TIEs, perceptions of vocabulary learning or attitudes to testing. Given that the participants' awareness of their own task approach is relatively low, interviewing offered the chance to rephrase and explain questions and gain more in-depth information.

To avoid researcher bias in the interviews, the same set of questions was used as a basis for all interviewees. However, digressing and further probing for information is inevitable and desirable. In addition to building a good rapport with the interviewees, it was important to formulate questions thoughtfully and in a way that did not encourage the interviewees to give 'convenient/pleasurable' responses. Brenner (2006) found that: "because qualitative interviews are based on a personal interaction, who a researcher is and how informants view the researcher are likely to influence the kind of information received in an interview." (p. 368). To avoid this 'halo effect'[17], I decided to inform the interviewees about the aim of my research. I also made it obvious that there were no 'right' or 'wrong' answers.

Another reason for a mixed-methods design is that, just as classroom learning, the processes involved in vocabulary acquisition are extremely complex and it is advisable to explore them from several perspectives. Using just one method would simplify or distort the picture, as different types of vocabulary test provide different impressions of vocabulary knowledge. Accordingly, it was necessary to indicate what type of knowledge could be inferred from the chosen test type and what type of knowledge could not be measured. Combining different measures achieves a more holistic insight. Whatever test format is chosen, it is important to distinguish between different word knowledge aspects (form, meaning, pronunciation, grammar) and different forms of mastery (receptive, productive). At the same time, the interaction between these two concepts must be accounted for; for instance when considering the results, both dimensions have to be discussed in connection (Schmitt, 2010).

Finally, it is necessary to situate my study on the experimental versus non-experimental continuum. Experimental studies can deliver a clear picture of cause-effect relationships through a comparison between an 'experimental' (intervention) and a control group (no

[17] Halo effect "refers to what happens when interviewees pick up cues from the researcher related to what they think the researcher wants them to say, thus potentially influencing their responses" (Mackey & Gass, 2005:174).

intervention). In a more natural set-up it is often difficult to disentangle the interferences of various other related factors. While a purely experimental set-up may be desirable, in an educational setting random sampling is often not feasible and researchers are often dependent on the "contexts that already exist" (Gass, 2010:11). Using a quasi-experimental design and working with non-equivalent intact class groups, "cluster random sampling" (Mackey & Gass, 2005:120) - has become an accepted research methodology in studies where randomization is impossible or impractical and is now the most common method in education (Dörnyei, 2007).

My study utilized a quasi-experimental set-up, in which investigations took place in authentic learning environments. The interviews were conducted on a voluntary basis. The data from interview participants was removed from the overall data pool of vocabulary testing, because these students were particularly aware of the focus of the study and discussed some of the target words in detail during the interviews. Keeping them as part of the overall data pool would have distorted the results.

Details of the data collection procedure are explained in the following section.

3.3. Procedure, Setting And Participants

First a two-part pre-test was conducted. Pre-test I was run to find the most suitable texts, target words, and pre- and post-reading exercises and to design the most suitable test. Pre-test part II was used to check the suitability of the chosen texts and target words. In the pilot study, the overall research design was tested and test-interviews were carried out.

The main data collection consisted of three sessions of in-class reading tasks and vocabulary testing and subsequent voluntary retrospective interviews with 1-2 students from each class. Over a three-week period, the learners worked on three reading tasks with different enhancement types, or as part of the control group, which read an unenhanced text. They read the text, answered the reading comprehension questions, and worked through the post-reading exercises. Immediately afterwards, they were asked to complete the vocabulary post-test (VPT). To allow for within group comparison, all participants worked through all three texts, but with different TIE types and were tested three times. I decided to work with a repeated reading/testing procedure, i.e. a scenario where the same students were tested several times, to be able to compare the behaviour of the same group of people with different types of TIEs (Eckerth & Tavakoli, 2012). To minimise the effects of repeated testing (test fatigue, test awareness), I split up and first analysed the data collected after the first reading/testing session and then the data gathered in the subsequent sessions.

Four weeks later a delayed vocabulary post-test was conducted. In this way, the reading/testing sessions appeared in the following way:

Table 3. Reading and Testing Sessions

	week/session 1	week/session 2	week/session 3	week 7
Class 1	Text 1 / Condition 1	Text 2 / Condition 2	Text 3 / Condition 3	
Class 2	Text 2 / Condition 1	Text 1 / Condition 3	Text 3 / Condition 2	
Class 3	Text 1 / Condition 2	Text 2 / Condition 3	Text 3 / Condition 1	
Class …	…	…	…	
Control group	unenhanced text	unenhanced text	unenhanced text	

Prior to data collection, I recognized that aspects like text difficulty and topic familiarity, might fetter reliability and validity of the research design. I also expected it as potentially problematic that the repeated reading/testing treatment could distort the findings. The respondents might become sensitized to the subject matter, which would influence the responses. To lessen these effects, a comparably large sample of EFL students representing a relatively homogenous level of proficiency was tested across groups and TIE-types.

I collected the data for my study in Danish secondary schools where I had easy access to teachers and students, and because this is the education system I am most familiar with. After initial contact with several schools and teachers was established, the first visit to the classes involved informal class observations and further conversations with the teachers. At a second meeting, the participating classes were informed about the research project and their role in it. A brief meeting with teachers and students before the data collection was used to answer questions and to collect ethical consent forms.

The chosen participants were students from twelve intact classes (269 participants overall) in schools, situated in and around Copenhagen. Each class consisted of approximately 25 students, who were aged between 16 and 19 years and attended the ‘2g’ class – the second year at Danish secondary school. They all studied advanced level English.

I used to teach at one of the participating schools, but was not acquainted with any of the research participants. However, because of this professional experience, I am familiar with the Danish education system, where English has a prominent place in the curriculum. English as a foreign language is part of compulsory education and is taught from around the age of 10. In secondary schools, depending on their chosen strand of study, students learn EFL for either 3 or 5 hours per week at ‘A’ (higher) or ‘B’ (standard) level. Over the last decades, Denmark has become more and more multicultural. As immigration and thus language diversity is mainly concentrated in Copenhagen, in classrooms there many bi- and trilingual students with many different language backgrounds can be found.

The language situation in Denmark, being a small country with only about 5 million native speakers, is worth mentioning. Because of the small number of people who speak Danish as a mother tongue, English is used widely in communication with most foreigners, often even with their Swedish and Norwegian neighbours, whose languages are closely related to Danish. Most Danes are exposed to foreign languages, and English in particular, from an early age. Therefore, Danes of all walks of life have a good oral command of English.

The participating students were all learners of English as a foreign language and had studied the language for 9-10 years in school. Their language competence can be compared to level B2 / C1 of the Common European Framework of Reference for Languages, i.e. they can be categorized as of upper intermediate proficiency. Testing learners of a lower level of proficiency would have made the procedure of finding the target words easier: elementary/lower intermediate learners are more uniform in their (lack of) vocabulary knowledge, so that it would have been easier to find words unknown to all of them. However, I had no access to this group of learners. Moreover, previous studies all worked with university students, i.e. learners at a higher level of proficiency. To make the results of my study comparable to previous research, the level of proficiency of my participants had to be at a similar level. The selection of the students was not based on gender or first language background. Rather, English subject leaders were contacted randomly and asked whether they were interested in participating in the research project. These teachers discussed with their students whether they wanted to learn more about the study so that they could decide whether they wanted to take part.

This random sampling approach in secondary schools seemed suitable to ensure that the participants form a representative sample. Mackey and Gass (2005:141) had pointed out that “randomization can enhance the experimental validity of a study. However, there are situations when randomization of individuals may not be feasible”. In order to not interrupt the classroom situation more than absolutely necessary, it was inevitable to work with intact classes. If intact classes are used, however, the researcher should carefully consider how the classes are assigned to treatment groups (Dörnyei, 2007). Several efforts were undertaken to ensure representativeness and generalizability.

3.4. Ethical Issues

This research study bore minimal risks for the participants[18]. Taking part in the treatment could rather be seen as a benefit, because of the additional language practice. The text topics were related to what had been studied in class and therefore served as good preparation for the students' exams.

However, there were ethical concerns to be taken into account. The reading tasks were designed for a research purpose and were not state-of-the-art teaching methodology. For students in Denmark, who are used to working with innovative teaching methodology and in very modern learning environments, this meant asking for patience. The participants were given the opportunity to discuss the purpose and had time to sign 'informed consent' forms or to opt out.

Informed consent requires that subjects should be provided with the "opportunity to choose what shall or shall not happen to them" and can only be guaranteed, if

> (1) sufficient information is supplied
> (2) subjects understand their role in the research, and
> (3) participation is fully voluntary (Mackey & Gass, 2005:27).

With regards to 1, I could not fully disclose all information about the study. Giving away the precise goals of my study ('investigating vocabulary learning from reading') would inevitably have made the participants aware of the focus and alerted them to the vocabulary in the texts. This would have counteracted the intention to explore vocabulary gains from *incidental versus intentional* reading. The research design therefore required that I revealed the purpose of the study in more general terms, i.e. 'What can be learned from reading in the foreign language classroom?', so that the participants had necessary, but not too detailed information. This procedure seemed acceptable as there were no risks for the subjects connected with it, and because the participants would be debriefed about the details of the study after the treatment (Mackey & Gass, 2005; Strike, 2006).

In relation to 2, all participants could ask questions in their L1 or L2 and were given enough time to read the information, which was phrased according to their level of proficiency. All the details were presented to the students orally in the first place, as it is widely advised (Mackey & Gass, 2005:31). The participants had the chance to discuss the research in either their L1 or L2 with their teacher, with the researcher, and with each other. A student information sheet

[18] According to the 2011 BERA guidelines, http://content.yudu.com/Library/A2xnp5/Bera/resources/index.htm?referrerUrl=http://free.yudu.com/item/details/2023387/Bera, accessed 28/9, 2015.

and a consent form was given to all participants of the main study. A specific student information sheet and consent form was given to all participants of the interviews. Thus, the participants had the chance to read about the research in their own time and possibly amend their decision about whether they wanted to participate.

Concerning point 3, it is important to point out that Danish high school students administer their own time in school. If they had decided not to participate in the study or leaving the experiment, they would not have been penalised for absence. Furthermore, all students had sufficient time to make a reflective decision (Strike, 2006) and were made aware that they had the right to withdraw from the study at any time without giving reasons. They knew that all of their data would be handled anonymously and would be used for research purposes only. It was made very clear that their participation was in no way related to class performance and that their teacher would not have access to any of their data.

Lastly, retrospective student interviews require some ethical considerations. Depending on the number of volunteers, 1-2 students from each class were interviewed (more if more were interested). The interviews took about 20 minutes and were conducted as soon as possible after the reading treatment and vocabulary testing. The interviews took place in the participants' school, i.e. in familiar and safe surroundings. In the interviews, students were asked to comment on some of the words they were tested about in the vocabulary post-test and on the way they worked with the reading task. These questions did not test their competence and were not supposed to put them under pressure to 'perform' well. It was made clear that participation in the interviews was not related to their performance in their English lessons and that their teachers did not have access to any related information. In addition, the students were assured that their participation was voluntary, that they could abandon the interview whenever they wished and that their data would be used anonymously. The interviews were conducted in English or Danish, depending on which language the participant felt most comfortable with.

3.5. Research Instruments

Boring texts or unsuitable target words would have been counterproductive to achieving the research goals. I had to ensure that the chosen text would not be uninteresting of too easy/difficult for the students. Therefore, clear categories were needed to select the words and texts.

3.5.1. Choosing The Texts

The following categories were considered for choosing the texts:

- text length,
- authenticity,
- text difficulty
- relevance and topic familiarity

Text Length

Depending on the overall complexity and familiarity of the topics, the texts varied slightly in terms of word number, but an approximate text length of 800 words was aimed for. This was a suitable length in relation to the available lesson time. The feasibility of this word count was measured and confirmed in the pilot study and by the teachers.

Authenticity

Text 'authenticity' is important in two different ways: The text type had to correspond with what the participants would perceive as a typical foreign language classroom text, as the study wanted to capture 'normal' classroom behaviour (e.g. concerning task approach). Therefore, 'authentic' is used here in the sense of 'authentic within the classroom'. Accordingly, the texts were chosen based on my professional experience and based on the way the teachers usually chose texts for their English lessons.

The other issue was whether the texts were 'genuine'. They had to be "representative samples of the kind of texts foreign language students will eventually encounter in the target country" (Gettrich-Ludgate & Tovar, 1987:80). Despite some necessary changes, it was important to make sure that the text kept its original form and message. Here authenticity concerns the changes made to the texts to make them suitable for the goals of the study, and is therefore closely related to 'text difficulty' (see below).

The print medium was chosen even though computers and reading from screen are an integral part of most Danish classrooms nowadays. However, at the start of this project, reading from paper was still the most common form and the easiest to operate in this project.

Text Difficulty

Text difficulty is not easy to determine, as it not only depends on text-internal, but also on several contextual and learner-related issues, and as it in turn influences overall task difficulty. Révész and Bruntfaut (2013) identified lexical complexity (e.g. lexical diversity and density[19]) as a factor that significantly predicts overall task difficulty. They found that "texts that contain a wider variety of words are presumably more difficult to process because they require the decoding of a larger number of unique words in the same amount of time" (p. 38). The study also confirmed their assumption that text difficulty correlates strongly with "both learner perceptions of overall task difficulty and the extent to which learners felt they completed the task successfully", i.e. that text difficulty is linked to learner motivation (p. 60). Therefore, and even though teachers would not normally go through such a meticulous screening process when choosing texts for their lessons, it was important for my research to find texts that were in terms of their lexical complexity appropriate for the participants' proficiency level. Text difficulty here relates to topic familiarity, lexical density and text coverage, i.e. how many of the texts' words were known by the readers. I therefore chose to combine different perspectives in order to come to a holistic understanding of the texts' difficulty level.

While topic familiarity was confirmed by the teachers, ensuring suitable levels of lexical density and text coverage was the greater challenge. 'Coverage rate' is the percentage of the vocabulary that is known by the reader (Waring & Nation, 2004). It was important that potential texts had to offer an appropriate balance of known versus unknown words so that the likelihood of sufficient comprehension would be maximized (Waring & Nation, 2004). There are several readability measures that capture a text's difficulty for a (foreign) language learner (e.g. the Fry Graph[20] or the Flesch-Kincaid). They usually put the number of words and/or syllables in a sentence or text in relation. I used the Flesch-Kincaid Grade Level, a formula commonly used in related studies (Azari, 2012; Fraser, 1999; Ko, 2012; Lee, 2007). While Text 1 ('Horror') scored 8.8 points, Text 2 ('Equality') scored 10.7, and Text 3 ('Divide') 10.6. Even though several teachers had assessed the three texts to be of the same difficulty level, according to the Flesch-Kincaid Level measurement, the results for Text 1 were lower. Still, I decided to include Text 1 into the study, because it seemed sensible to follow the teacher's recommendations. The chosen texts had to be suitable not only regarding difficulty, but also with respect to topic, curriculum and learning situation. I trusted that the teachers knew their students and their abilities sufficiently well to judge the texts' suitability for the students. When I analysed my data later, I found that text difficulty sometimes had a significant effect on my

[19] The proportion of content words to the total number of words in a passage.
[20] http://www.readabilityformulas.com/fry-graph-readability-formula.php, accessed 14/8, 2009

findings. In hindsight, I therefore think that substituting Text 1 with another text, which more precisely matched the other texts' level, would have made my quantitative results clearer.

Two other questions were important; one was whether the learners knew the target words within the texts. Here, the role of context had to be considered as words might be known within context, but not in a decontextualized form. Therefore, the pre-test consisted of two parts, with the potential target words occurring in decontextualized form (part I) and in contextualized form (part II).

The second question was how many words exactly learners need to know for effective L2 reading. The precise numbers are not fully clear. While earlier research assumed that at least 95% of the words need to be known (Laufer, 1989; Liu & Nation, 1985), it is now believed by some that at least 98% are needed (Hu & Nation, 2000; Schmitt, 2010). It is difficult to measure and compare results, because there is no common understanding of what counts as a 'known' word. Still, to be able to understand such high percentages of words in texts, it is likely that learners have to be familiar with up to 6,000-7,000 word families for spoken discourse and 8,000 – 9,000 for written (Nation, 2006).

The specific numbers, may depend on the type of text (academic and technical texts need higher coverage rates to be understood). In this study, appropriate text coverage was ensured by checking the knowledge of a high number of (according to the British National Corpus, BNC) lower-frequency words in the pre-tests. Non-target words that were unknown by most participants (90 %), but were important for text understanding were substituted with more common words.

Text coverage is also important because it may affect reading strategies. Research suggests that, unless the reading is done at a high level of vocabulary coverage, little learning will take place when guessing from context, which is a very common strategy (Waring & Nation, 2004). To ensure a suitable language level, it was necessary to substitute low-frequency non-target words by more common synonyms, while keeping in mind the issue of authenticity. Then the texts were shortened to approximately 800 words by cutting out those phrases and sentences, which did not significantly contribute to the overall message of the text. Thus, it was less likely that learners would be distracted by difficult and possibly salient unknown vocabulary and could focus on the reading process. Finally, the enhancements were added.

Text difficulty also depends on the reading situation. Reading in a classroom is necessarily more formalized and puts learners under more pressure than reading for pleasure. To be able to take all the factors described above into consideration, it was necessary to trust the teachers' expertise and experience.

Relevance and Topic Familiarity

Previous studies suggested that a greater level of topic familiarity and knowledge of a given subject help paying attention to input during reading, enabling better comprehension and, in turn, superior memory performance (Ellis, 2001; Nassaji, 2002; Pulido, 2003). This indicates that different levels of topic familiarity might influence the comprehension of the text as well as the acquisition of vocabulary. Lee (2007) highlighted the importance of topic familiarity for freeing attentional resources to focus dually, on form and meaning. Therefore, I tried to find texts that as much as possible resembled the topics and the type of text the students usually work with. First, I consulted their teachers. They usually use texts from textbooks, magazines, or the internet. To follow their method and to minimize the probability that the participants might be familiar with any of the texts used in the study, I used English teaching websites, newspapers and popular scientific magazines for the search. The selected texts cover topics that are relevant for topics dealt with in their English lessons as well as of general interest (from science, popular culture). After examining various texts according to these criteria, they were finally narrowed down to seven texts.

The objective was to find texts the students would be interested in. Therefore, finally, and after having checked the texts for all categories explained here, as part of pre-test II, the participants were asked to report on in how far the topics of the texts had been interesting to them (Fig. 2).

Figure 1. Pilot Study Text Selection Questions

Finally,
… please indicate whether you found the text topic 'interesting' by circling the appropriate number:

1	2	3	4	5
very interesting: I found the text very interesting, and find this topic very relevant and/or entertaining	*interesting*: this was quite interesting to read and it is a topic that I find rather important	*rather interesting*: it was neither interesting nor boring to read this text and to think about this topic	*not very interesting*: I didn't find the text very interesting or entertaining and don't think this is very relevant	*not at all interesting*: this is a boring text and topic and I don't think this topic is at all important

Based on these accounts and further consultation with the teachers the three texts 'Divide'[21], 'Horror'[22] and 'Equality'[23] were chosen[24].

3.5.2. Choosing The Target Words

When choosing the target words, it was particularly challenging to find a balance between unknown and at the same time relevant words. I investigated how previous studies had dealt with this problem. In their empirical Involvement Load study, Hulstijn and Laufer (2001) had chosen ten low-frequency target words. Possible knowledge of the words was assessed through a list in which the participants were asked to translate or explain the target words. This test was conducted with groups of learners at a similar proficiency level as the groups participating in the experiment proper. Target words had to have an average mean of knowledge lower than 1 out of 10. Pre-knowledge of the target words in the experimental groups was checked at the end of the experiment. Laufer and Hulstijn, however, did not report on how they dealt with semantically opaque words, or the matter of contextualised versus decontextualized occurrence of these words. Keating (2008) circumvented these problems by choosing eight nonsense words. Even though this is an effective approach, it is not an authentic one and was therefore not deemed suitable for a classroom study. To find the ten target words, Kim (2008) asked two language students to find unknown words from a list. The target words were chosen among the words unknown to the students and finally the teachers were asked to confirm that the chosen words would really be unknown to the participants. However, consulting only two students and relying on the teachers' judgement alone did not seem a reliable procedure. Therefore, this approach was not pursued either, but the target words were chosen through a 2-part pre-test. In part I, all low-frequency words were presented in decontextualized form in a list and the participants were asked to report (the degree of) their knowledge. In part II the remaining potential target words were presented in context.

The target words had to be suitable in relation to four categories:

- frequency,
- familiarity,
- relevance,
- clarity and difficulty[25].

[21] 'Divide' text: about the remaining social division between black and white Americans (Orlando Patterson, NEWSWEEK magazine, Nov 10, 2008.
[22] 'Horror' text: about, for instance, vampires and zombies in horror literature (Joseph VanBuren, Psychology, December 20, 2007; http://socyberty.com/psychology/the-psychology-of-horror/ accessed 16th July 2008.
[23] 'Equality' text: about the changing role of men and women in marriage (Stephanie Coontz, The TIMES, April 11, 2007.
[24] For the three chosen texts, the reading comprehension questions and post-reading reflection activities, please see Figures Appendix 1, 2 and 3.
[25] The way these categories as listed here does not represent an order of importance, but rather shows the principles that had to be balanced when the target vocabulary was chosen.

Frequency

The two main guiding principles were that the target words had to be unknown and very relevant to the reading task. To ensure the first, the potential target words were narrowed down the by checking their 'frequency of occurrence' in a language corpus. I used the BNC and the American Corpus. Frequency carries several implications relevant for vocabulary acquisition. It is assumed that the most frequent words are usually the most useful ones because frequent words have greater chances of being met and used (Shin & Nation, 2008). In this study, low-frequency words were chosen, as they were more likely to be unknown to the participants, in order to heighten the study's validity. The other reason arose from language teaching reality and the natural distribution of words in the English language. The most common words (the 1,000-3,000 most common words, according to Nation, 2001) are usually considered as so important that they are taught explicitly. The large number of 'mid-frequency' words, the 3,000 – 9,000 level, however, is problematic as these words are relevant, but so numerous that they cannot all be dealt with in the classroom. This lower frequency vocabulary is relatively scarce in many texts (Cobb, 2007; Pellicer-Sánchez & Schmitt, 2010). Nation (2006) found that only 4.90% of the tokens in the BNC come from the 4,000-9,000 frequency levels, but make up 25.25% of the types. Pellicer-Sánchez and Schmitt (2010) highlighted the importance of this circumstance for reading research: "This scarcity of mid-frequency vocabulary has serious implications for incidental learning", namely that a great part of this vocabulary has to be learned incidentally (p. 32). For my participants, this range of vocabulary was particularly relevant, as they were of upper intermediate proficiency level. Therefore, the target words were chosen from this frequency level.

Frequency was excluded as a factor in my study, although I had initially planned to include it. However, while developing the research set-up, it became obvious that the texts were too short to repeat target words often enough for measurable effect[26]. Forcing the target words several times into the texts would have been artificial. In addition to this, including frequency as a variable, would have overloaded the already complicated research design.

Familiarity

Ideally, all target words would be unknown to all participants. Words which turned out to be known according to pre-test part I were deleted from the list. If more than 20% of the

[26] Previous research had shown relatively univocally that words have to be repeated many times in order to lead to a measurable effect (Eckerth & Tavakoli, 2012; Horst, et al., 1998; Rott, 2007, Vermeer, 2001).

translations/example sentences given by the participants in pre-test part I were correct, the word had to be excluded from the list of potential target words[27]. The remaining words had to be unknown to most participants of pre-test part I. If more than two of the 59 pre-test participants (3 %) provided either a correct translation or a correct sentence containing the word, it was deleted from the target word list. To avoid target words that could be easily inferred from Danish (e.g. cognates), the potential English target words were compared to their Danish counterpart. This comparison of Danish and English words was discussed with native speakers and the teachers, all of which have a very good command of both English and Danish.

Finally, the words were tested in both parts of the pre-test: Even though all low-frequency words were included in pre-test part I, some words with higher frequency emerged as potential target words in pre-test part II. Thus, the two tests complemented each other.

Relevance

According to the noticing hypothesis, a word needs to be noticed to be learned. Therefore, the potential target words had to be salient enough to be noticed as *unknown* in context. This was tested in pre-test II, where the target words selected based on the findings from pre-test part I were checked for their familiarity (section 3.7).

Just as importantly, however, the target words needed to be relevant. Similar to Peters et al. (2009), a relevant word was defined as a word that was closely connected to key ideas of the text, i.e. was essential to answer a reading comprehension question. Peters (2007b) found that word relevance had an effect on immediate and delayed recall of word meanings, albeit, as shown in a later study, not on recognition (Peters et al., 2009). Therefore, it was important for my study that the meaning of the word had to be relevant within the text, in the sense that the sentence could not be easily understood without knowledge of the word. Through this forced attention the target words would be embedded in the text naturally and would become an almost inevitable focus in the reading process. Several of my fellow PhD students confirmed the importance of the target words in the texts and the teachers' judgement was very helpful, too.

I also wanted to find words that were relevant for the students to learn, i.e. words that they might actually encounter in classroom texts. While this concern of authenticity was less relevant for the choice of target words, it was taken into account in the choice of texts (section

[27] This number seems sensible, because there were only very few translations/sentences given and seemingly only by the participants most confident of their knowledge. Thus, this is all in all a very low percentage.

3.5.1). Therefore, primarily words that originally occurred in the chosen texts were considered and substituted with a lower frequency substitute, if necessary.

Clarity and Difficulty

Different items have different saliency values, which influences how well they are suitable for TIEs (Cho, 2010; Han et al., 2008). Therefore, avoiding semantic opaqueness and form-confusion was important (Laufer, 2012). Words with similar forms, so called 'synforms' (Laufer, 1988) can be confusing as their transparency is deceptive, and some other words have very diverse meanings, which might even appear within a single text. To avoid confusion, the meanings of the target words had to be constant throughout the text. The target words were evenly distributed regarding word class: each text contained 3-4 unknown verbs, nouns, and adjectives. The target words also varied in length. It was difficult to further control for intrinsic word difficulty (e.g. spelling, pronunciation, Laufer, 2012). However, each text contained a target word, which was opaque (Text 1: staple, Text 2: sphere, Text 3: to rear). In this way, it was possible to examine the participants' test taking behaviour, for instance guessing.

To avoid confusion, difficult idioms like 'to curry favour'[28] or similar conspicuous, low-frequency words were eliminated. From the remaining target words those were chosen, which best fitted into the texts, i.e. those which were already part of the original text and those which least interfered with the reading flow and text understanding. For example, in the 'Equality' text, 'whim' was kept as a target word, but the verb 'to indulge' had to be substituted by 'to tolerate'. Finally, the following target words were chosen:

Table 4. The Chosen Target Words

Text 1 **'Horror'**	**Text 2** **'Divide'**	**Text 3** **'Equality'**
1. encounter	11. entail	21. deem
2. notion	12. besetting	22. rigid
3. glimpse	13. suffrage	23. mould
4. alleviate	14. ascent	24. whim
5. innate	15. incarceration	25. pollster
6. staple	16. emerge	26. disparate
7. grasp	17. profound	27. impose
8. ravenous	18. poised	28. spouse
9. eerie	19. sphere	29. ubiquitous
10. invoke	20. persist	30. rear

These target words appeared enhanced in the texts together with five enhanced distractor words, to diminish the training effect. To eliminate the effects of other factors like topic

[28] Originally in the '*Equality'* text.

familiarity and interestingness, the treatments varied across groups, i.e. all participants read all three texts but with varying enhancement types or without any enhancements (control group). Moreover, the order of the texts varied across groups to avoid distortion of results through a 'training effect'.

3.5.3. Choosing The Enhancements

Previous research has investigated the effects of L1 versus L2 enhancements (e.g. L1 versus L2 – Bell & LeBlanc, 2000; Ko, 2004; Jacobs, 1994; L1 - Laufer & Goldstein, 2004; L2 - Pellicer-Sánchez & Schmitt, 2010). Often, differences between different types of TIE were found primarily regarding the learners' preferences, but sometimes also with respect to their effect on vocabulary acquisition or comprehension. As explained in section 2.6.1, several studies found no difference in their influence (Jacobs, 1994; Ko, 2004), others found L1 enhancements to be of superior effect for vocabulary acquisition (e.g. Laufer & Shmueli, 1997; Yoshii, 2006). I decided to use L2 enhancements in order to reflect what most EFL learners in Denmark are used to from their language learning materials. The teachers confirmed that this was an appropriate choice, considering the participants' proficiency level. This is in line with research findings.

The three reading conditions (= TIE-types) and the unenhanced control condition represent different types / degrees of enhancement and were coded accordingly:

- control group = C0 (unenhanced)
- Condition 1 = C1 (bold printed target words)
- Condition 2 = C2 (glossed target words)
- Condition 3 = C3 (target words bold printed and glossed)

The control group was allocated the number '0' as the participants' text was completely unenhanced. Number '1' was given to Condition 1 in which the target words were bold printed. Condition 2 (L2 glossed target words) was registered as '2'. Finally, Condition 3, i.e. bold printing and L2 glossing the target words, was coded as Condition '3'.

These enhancement types were chosen to replicate previous related studies (e.g. typographical enhancements like italics, underlining and bold-printing were for example used by Barcroft, 2003; Simard, 2009; Wong, 2003; glossing was used for instance in Izumi, 2003; Leow, 2001; Min, 2008) and because they represent the most common enhancement-types

used in Danish EFL secondary schools textbooks[29] and are therefore familiar for the students. Studying the effects of these typical enhancement forms seemed most suitable, as one of the aims of this project was to investigate real-life learning behaviour.

As different types of enhancement may have a distinct impact on vocabulary acquisition, they are regarded as distinct "reading conditions" (e.g. Barcroft, 2009; Watanabe, 1997; Yoshida, 2008). Through using bold-printing and glossing (and the combination of these) I hoped that a comparison between the effects of different learner foci would be possible.

Condition 1 raises salience through bold-printing, a visual form of highlighting, and guides learner attention to word form (Bruton, 2011). This may affect reading and vocabulary acquisition. Through the typographic highlights, the reading flow may or may not be interrupted. While they do not encourage learners to look up and read a definition in the margin, typographic enhancements may interrupt, as learners receive no feedback to any hypotheses they might be forming while processing word information. Glossing, Condition 2, is likely to guide attention to meaning and provides the reader with immediate feedback through vocabulary help. At the same time, it requires readers to interrupt the reading flow, to decontextualize the word, read the definition in the margin and then process this information to continue reading. Condition 3, where target words are both bold-printed and glossed, potentially guides attention from word form to the explanations given in the margin to make a focussed process of reading and checking of both form and meaning possible. Bruton et al. (2011) suggested that examining different combinations of influences (e.g., bolding plus marginal gloss versus bolding/no bolding plus linked gloss) was an important research need.

Through glosses, the text is manipulated typographically as well, but what seems more important from the point of view of vocabulary learning, is that learners are given lexical help through immediate 'feedback' on their hypotheses on the word's meaning or else, they are helped along with the reading process (Færch et al., 1984). This feedback contributes to explicit language learning; which is believed to entail the formation of hypotheses about the target language based on linguistic input and the learner then testing these hypotheses (Haastrup, 1991). In the case of vocabulary acquisition this may for example happen in subsequent encounters with the vocabulary item, for instance by producing output ('output hypothesis', Swain, 2005), or by looking up the word in a dictionary (Færch et al., 1984). This testing can lead to the rejection of a language hypothesis if the learner interprets the feedback as being in contradiction to the hypothesis. Alternatively, if supportive of his hypothesis, a new rule about the L2 may be learned, for instance the meaning of a new word. This is a simplified

[29] For instance Damskier & Weidick (2003); Engberg-Pedersen et al., (2004); Ramberg-Beyer (2013)

view of vocabulary acquisition as erroneous interlanguage is much more difficult to rectify. The confirmation or rejection of formed language hypotheses is not sufficient, the vocabulary acquisition process is more intricate and many facets of knowledge are acquired implicitly. Still, the concept of feedback must be considered when investigating glossing, on the one hand because it highlights the crucial role of input for vocabulary acquisition and on the other hand because glosses intervene with these acquisition processes, positively or negatively.

The definitions for the target words glossed in the margin were found in the online version of the Longman Dictionary of Contemporary English[30] and with the help of the BNC. It was essential that these were short, comprehensible, and represented the specific meaning of the word within the text.

3.6. Reading Tasks And Follow-Up Exercises

The task treatment consisted of three parts; text reading, answering comprehension questions, and completing follow-up exercises. To measure the retention rates for word learning from reading texts with different forms of enhancements, the learners were asked to participate in another 'exercise'. They were not informed about the nature of this 'exercise'. This was to avoid them focussing on the vocabulary. The instructions and goals for all these task components are explained below.

3.6.1. Reading Task Instructions

The reading tasks were handed to the students in a stapled set of printed text. On the cover sheet, the instructions for the three reading tasks were the same except for the TIE type. These instructions were as follows:

Figure 2. Reading Task Instructions - Examples Taken from 'Horror' Text

INSTRUCTIONS

Reading Task

This article explores why so many readers are attracted to horror literature and explains different aspects of the concept of fear. Please read the questions on the next page. Reading the text will then help you to find the answers.

To help you with the reading, difficult words are in bold type and are explained in the margin**.**

margin**: empty space at the side of the page**

[30] http://www.ldoceonline.com/

2. Reflection Activity
After having answered the reading comprehension questions, please do the next activity. There you have to think more about fear and horror literature.

3.Matching Exercises
Finally you will have to complete some general language tasks. There are four exercises in this part. Please follow the instructions carefully.

Thank you for participating, I hope you enjoy the reading!

These instructions were followed by comprehension questions on the second page of the set. In order to reflect classroom reality no instructions related to the enhancements were given. In Danish textbooks enhancements are usually provided, but with no specific instructions linked to them. Moreover, I did not want to alert the students to the purpose of my study.

3.6.2. Reading Comprehension Questions

As explained in the Literature Review, comprehension is closely linked to vocabulary acquisition from reading. Peters et al. (2009) assumed that reading comprehension questions would "prompt L2 learners not only to look up (relevant) words […] but also to process the lexical information more elaborately in order to provide a correct answer to the comprehension questions (p. 117). Similarly, de la Fuente (2006) found that in pre-planned and guided tasks, target words were better incorporated than in free unguided activities. I used reading comprehension questions to make sure the students engaged in normal reading behaviour, meaning that students would try to understand the text as a whole and thereby not solely focus on the target words. However, it was essential to write comprehension question that related very narrowly to these target words, so that these would necessarily have to be processed.

To check comprehension, multiple-choice questions are frequently used in foreign language classrooms and in reading studies, as they are comparably easy to understand and to analyse. Another option are true/false questions. Both types were used in my study. These types of questions were chosen to follow previous studies (e.g. Hulstijn & Laufer, 2001; Keating, 2008; Kim, 2008). As a slightly more open approach, I added a third type of question, in which I asked the readers to answer questions about the text in their own words. This variety of question types includes open and closed approaches, thus avoiding some of the shortcomings of multiple-choice questions (lack of creativity, guessing instead of thinking). However, as comprehension was not directly assessed and analysed in my study, these considerations were a practical, rather than of a research-related concern.

For the wording of the multiple-choice questions general points of interest related to the text-topics were considered. It was also important that the items covered explicit and implicit

information. By following the technique described in Hu and Nation (2000), a native speaker helped me to divide the text into idea units, and the questions were allocated accordingly. For each question, four answer options were provided. These had to be close to the text, not too obvious or too far-fetched, and capable of encouraging narrow reading but not close reading. After the pilot study, however, these options were reduced to three. This was done because it became obvious that three options were sufficient to serve the purpose in this study, to save time and because it was difficult to find four plausible options in an 800 word text without them becoming bizarre.

In Keating's (2008) study, the participants were not permitted to refer back to the text while answering the comprehension questions. By contrast, in my study using the text to answer the questions was permitted as it encourages active reading and reflects the real-life classroom situation.

3.6.3. Post-Reading Activities

The decision to integrate post-reading activities into the task design was based on two considerations. On the one hand, I wanted the tasks to be authentic. Working on exercises that expand on a text's topic is a natural part of L2 classroom reading. On the other hand, the tasks had to be appropriate for the research procedure. They also took the focus off the VPT. Time limitations were imposed on the reading activity to make sure that the tasks could be completed within the regular class period. Time on task was assumed to be equal for the three texts. This assumption was tested in the pre-tests.

To facilitate incidental/intentional vocabulary learning it was necessary to find meaningful tasks that focussed on the overall idea of the text rather than on language issues. The aim was to find open ended tasks which ask for a general opinion, rather than for statements that require narrow text reading. Different options for such activities were planned based on my professional experience. The teachers were consulted about the suitability of these tasks and with the help of their suggestions the tasks were adapted.

Finally, the post-reading exercises, which were neutrally called "reflection activities", were the following: The 'Horror' text (Text 1) was followed by a "reflection activity" called "How fearful are you?", which asked the students to tick off how scared they are in a Likert scale type response grid. For the 'Division' text (Text 2) a ranking task about immigration was chosen. For Text 3, ('Equality') students had to complete sentences which asked their opinion about gender equality. These tasks were all related to the issues raised in the texts, but did not lead the readers back to closely studying what they had read.

The reading comprehension questions and the post-reading activities were merely used to give an authentic design to the tasks and to distract from the vocabulary focus. However, it was not within the scope of this project to integrate them into the data analysis.

3.7. The Tests

Finding the most fitting test format is of importance for any vocabulary study and requires careful preparation. I consulted related research in the field, conducted two pre-tests and a pilot study, and analysed the data collected in them in relation to my research questions and concerns of test validity.

3.7.1. Pre-Tests: Determining The Target Words And Test Format

Two pre-tests were conducted to find the most suitable target words and test format. I had started the process by choosing potentially suitable texts (section 3.5) and considered possible target words in those. The goal of the first part of the pre-test was to find the most suitable target-words and test whether participants would know them when encountered in decontextualized form.

Testing the word knowledge of a group only representative of the 'experimental group', but not the main-study participants themselves, is a procedure which was chosen for similar studies (Hulstijn & Laufer, 2001; Kim 2008). However, this cannot be regarded as the ideal procedure. In the present study, various circumstances made it impossible to conduct the pre-test with the representatives of both the control and the experimental group, because the main data collection was carried out at the beginning of a new school year, with new classes, so that pre-testing these particular students in advance was not an option.

3.7.1.1. Pre-Test I: Objectives

For the pre-test, a design which would cover a wide spectrum of word knowledge was needed. The first part of the pre-test was run to find the most suitable texts, target words, and pre- and post-reading exercises and to design the most fitting test set-up. The instructions were the following:

Table 5. Pre-Test I: Instructions

Please fill in ALL correct boxes. If you know how to translate the word and also how to use it in a sentence, please do both. Thank you!

	English	**the word looks**				

#		somewhat familiar	I know the word	I know the word and can translate it into Danish	I know the word and can use it in an English sentence	I don't know the word
1	endure					
2	…					

This task is an adapted version of the various 'Vocabulary Knowledge Scales' (VKS). They have been used in many vocabulary acquisition studies (e.g. Hulstijn & Laufer, 2001; Joe, 1995, 1998; Kim, 2008; McNeill, 1996; Min, 2008, Scarcella & Zimmermann, 1998). This format relates to Paribakht and Wesche's (1996) VKS. Contrary to a simple translation test, these scales account for the incremental nature of word knowledge as they test different levels of vocabulary knowledge, starting from complete unfamiliarity. At the highest level, test takers are asked to use the target items in original sentence writing. As Kim (2008) points out, "the major benefit of the VKS is that it elicits students' perceived knowledge of vocabulary items and allows verification with demonstrated knowledge" (p. 300). While one of the shortcomings of this test format may be that these scales are not measuring many other relevant forms for word knowledge (e.g. the phonological dimension or semantic variety of word meaning, Bruton, 2009), and that sentence writing is student dependent and grammaticality is irrelevant, they elicit different degrees of active and passive word knowledge. Thus, they serve the purpose of this pre-test well, because it was not the aim to investigate vocabulary development. Therefore, I was not interested in 'full word knowledge', but rather in exploring the subjects' ability to demonstrate their knowledge of form-meaning relationships of the potential target words. The chosen pre-test format showed which of the potential target words were deemed unknown by the learners ('I don't know the word') and gave an insight in whether the learners were able to supply L1 equivalents of L2 words ('I know the word and can translate it into Danish'). Furthermore, it shows whether they also knew how the L2 word connected with other lexical items or grammatical structures ('I know the word and can use it in an English sentence'). The column 'The word looks somewhat familiar' elicited whether words were only known with regards to their form, while L1 meaning and L2 (grammatical) connections were possibly unknown. As the analysis of the pre-test data showed, information given in some of these columns overlapped and thus supported the notion of vocabulary knowledge developing along a continuum rather than within clear cut categories.

The particular order of the columns in the table was chosen to represent the above-mentioned spectrum of word knowledge: it ranges from an (assumed) familiarity with the word's form, via (assumed) knowledge of the word in either its L1 equivalent or L2 usage or both, to not knowing the word at all. The column 'I don't know the word' was put into final position to encourage students to carefully think about their word knowledge first and only as a last resource to tick off the 'unknown' option. Experience from using similar tests in another

vocabulary acquisition project (Eckerth & Tavakoli, 2012) has shown, that putting this column in primary position often leads less enthusiastic students to tick this box immediately without bothering to retrieve existing knowledge first. It is assumed that this order of the columns also corresponds to various levels of difficulty. Being able to use an L2 word in an original sentence is regarded to be more difficult, as this requires deeper processing than giving the L1 translation (Joe, 1995; Laufer, 2005).

3.7.1.2. Pre-Test I: Limitations

Despite the popularity of vocabulary knowledge scales, there has also been some criticism of this type of tests (Read, 2000; Stewart, 2011; 2012). They do not consider polysemy and do not provide information about test takers fluency in using the words. As the data collected from scales is nominal and not ordinal, interpreting the translation and writing scores can be problematic; there is also often insufficient guidance in relation to tasks, for instance 'Write a full sentence', when it is unclear what constitutes a full sentence (Waring, 2002).

Moreover, the chosen order of the different scale levels might have been confusing as they did not represent a clear and consistent increase of vocabulary knowledge. For instance, while "the word looks somewhat familiar" gauges partial knowledge, the difference between "I know the word", "I know the word and can translate it into Danish", and "I know the word and can use it in an English sentence" is not clear. It would have been preferable use mutally exclusive categories and to start with the highest level of knowledge ("I know the word and can use it in an English sentence") and then gradually go down the knowledge levels, ending with "the word looks somewhat familiar" (with "I don't know the word" in final position). This criticism was taken into account in this study in two ways: a second part was added to the pre-test design (section 3.7.1.4) and the sentence writing part was removed from the main-study vocabulary post-test design (section 3.7.2.1)

However, the VKS was chosen as it was sufficient for the purposes of the pre-test. This part of the pre-test was not primarily concerned with vocabulary development. Rather, the collected data was used to narrow down the number of potential texts and target-words. In combination with the second part of the pre-test, it was assumed, that the VKS would provide sufficient information about the participants' word knowledge prior to the reading treatment. Moreover, Waring (2002[31]) proposes that "instead of the researcher trying to pin down the subjects' knowledge into pre-set categories" to "have the subjects decide for themselves what their own knowledge is and report that". This suggestion was taken up in this study in the interviews, in which students reported details of their target word knowledge.

[31] http://www.robwaring.org/papers/various/scales.htm, accessed on 20/1, 2016.

3.7.1.3. Pre-Test I: Procedure

Before the participants were given the pre-test, they were informed about the purpose of the study, had the chance to ask questions, and the instructions were clarified. It was explained that filling in as many of the different columns as possible was important for the research as this represented different types of word. The participants were then given the list of potential target words. These were taken from the seven texts which had been judged to be potentially suitable for the reading tasks (section 3.5.1). The list consisted of 170 words and took participants approximately 40 minutes to complete; which was longer than anticipated. Contrary to their teachers' predictions, many students were familiar with a high number of the tested words and had to write many more translations and original sentences than expected.

To make this task easier, half of the students were invited to start working from the end of the list. This was to ensure that all target word items were covered, even if motivation should diminish while filling in the list. It had been necessary to include that many words in the pre-test to get a conclusive view of as many words as possible contained in the potential texts.

3.7.1.4. Pre-Test I: Analysis And Results

The pre-test results helped preparing the main study. Most participants did not show awareness of different parts-of-speech and translated nouns (e.g. *concern*) as verbs ('bekymret' - concerned) or nouns ('thrill') as adjectives ('begejstret' - thrilled). Many students simply used the more familiar word form, e.g. *urgent* instead of *urgency.* However, even if the translation was not fully correct in terms of part of speech, it was counted as a 'correct translation', as it was assumed that students knew the meaning of these words in context. While the successful students distinguished themselves by demonstrating precise knowledge of word meaning, generally the results revealed half-knowledge: for instance, students translated *gap* as 'hul', (eng. hole), demonstrating only partial knowledge of the word's meaning. There were also many cases of lack of knowledge. For instance in the translation of '*retail'* as 'daglig', (eng. 'daily') or *gap* as 'forskel', (eng. difference). These cases were counted as not-correct/unknown.

It became obvious that the categories 'I know the word' and 'The word looks somewhat familiar' were ambiguous. The participants seemed confused by them and in particular the latter category did not yield any valuable insights. This was taken into account in the later stages of this project, where these words were retested in pre-test II.

The results of three students were exceptionally better than those gathered from the other participants. After the teacher had confirmed that one of these three was a native speaker of English and that the other two were very successful participants in a pan-European EFL translation competition, their scores were removed from the pre-test data set, as their level of

language competence was not representative of the average student of that sample. After the pre-test, the texts were reduced to five and some potential target words were excluded because they were too well known.

As the target words were presented in a de-contextualized form, the test did not reveal whether subjects would know their meaning within a text (see for instance the difference between *entry* for 'door' versus *entry* in a dictionary). While this was not considered problematic in other studies (Hulstijn & Laufer, 2001, Keating, 2008, Kim, 2008) it was recognized as a limitation in the present study, and a second pre-test was therefore conducted.

As important as finding suitable target words, was designing a vocabulary post-test that would reliably measure the initial word knowledge gains from reading.

3.7.2. Vocabulay Assessment Types And Requirements Of The Test Format

When discussing vocabulary assessment, it is necessary to explore the various ways in which measurement formats differ. Read (2000) usefully distinguished three dimensions of vocabulary assessment, which represent use- and context-oriented ways of assessing vocabulary. He distinguished vocabulary knowledge tests with regards to whether they are discrete or embedded, selective or comprehensive, or whether they measure words in context or in decontextualized designs. These categories lend themselves to describe the two major traditions that have emerged in this field, and which treat vocabulary knowledge quite differently. One regards vocabulary as a distinct component of language, in which words can be treated as decontextualized individual items, mostly without considering, for instance, their grammatical behaviour (Laufer & Nation, 1995; Laufer & Yano, 2001; Meara, 1992). The other tradition on vocabulary testing underlines the importance of seeing vocabulary as inevitably linked to other language features; assessment therefore deals with words in context (Read, 2000; Read & Chapelle, 2001; Singleton, 1999).

In the first tradition, form-meaning type tests such as multiple-choice or translation tests that assess the lowest level of word knowledge, are very common (Waring & Nation, 2004). While these tests are handy to test large numbers of words, they have several shortcomings (e.g. test items appear decontextualized, work through a process of elimination, and can be solved fairly successfully by guessing correctly, for details see section 3.7.2.2). These unidimensional tests also "understate the importance of other types of word knowledge and possibly overstate the importance of the form-meaning relationship type of word learning when conducting incidental reading research" (Waring & Nation, 2004:16). Context-dependent vocabulary tests as used in the second tradition, however, are particularly time consuming and scoring can be ambigious.

As a way out, research has shown that for word learning studies, multiple measures lead to the most reliable results. Therefore, each target word should be tested in several different ways (Haastrup & Henriksen, 1998; Schmitt, 2010; Waring & Takaki, 2003; Webb, 2007; 2013) and discrete measures should not be seen as competing measures but as "measures tapping different strengths and aspects of vocabulary knowledge" (Nation, 2007:42). This is especially true when investigating incidental learning from reading, where research had found that "the use of several tests was necessary to gain a more accurate and balanced picture of learning. There is no one best way of testing learning. Each test reveals another facet of information about the kinds of learning that can take place" (Waring & Nation, 2004:17).

With regards to my study it was important to consider that vocabulary acquisition is an incremental and cumulative process, which can be imagined as moving back and forth on a knowledge continuum. One goal of my study was to explore this continuum with respect to which types of word knowledge are acquired when a word is encountered once in an enhanced form while reading with either an incidental or intentional approach to vocabulary acquisition. It was therefore necessary to find test formats that would measure different facets of word knowledge, ranging from the lower end of the continuum to more advanced knowledge. Thus, the test had to be sensitive to small gains and partial word knowledge, as it was unlikely that high-level word knowledge, such as knowledge about how a word behaves grammatically, could be achieved from encountering a target word once during reading. This had previously been suggested by Swanborn and DeGlopper (1999), who claimed that one could not "expect students to learn a dictionary-like meaning of a word after only one encounter" (p. 278). It is likely that most learners would display types of knowledge located at the lower end of the knowledge continuum, such as knowledge about basic form-meaning connections, the "essential 'core' knowledge" of a word (Schmitt, 2010:153). This level of competence can be tested in receptive test formats.

The components of the chosen vocabulary post-test are explained in the following section.

3.7.2.1. Measuring Vocabulary Gains – Adapting the CATSS-Test

To measure vocabulary gains I used the monolingual version of the 'CATSS' test (Computer Adaptive Test of Size and Strength; http://catss.ga/; the bilingual version of the test was developed in Laufer & Goldstein, 2004) and adapted it to the specific purposes of my study. The test was developed by Laufer et al. (2004) for the purpose of diagnosing interlanguage development and specifically to measure the strength of the acquired form-meaning link. This test was chosen because it fulfils many of the requirements outlined above.

Many aspects of word knowledge complexity are captured by the CATSS set-up, as it is a test of both vocabulary size and strength[32] and has the advantage of providing an insight into knowledge of word meaning from different perspectives and partial word knowledge is also revealed. The CATSS test is based on the three assumptions: 1. that the form-meaning link is the "most important component of word knowledge", 2. that there are different degrees to knowledge of meaning, and 3. that knowing the meaning of many words is more important than knowing a few words in depth (Laufer et al., 2004:209). Laufer et al. distinguished four 'modalities' of knowledge of meaning. They understand the ability to retrieve word form as `active/productive' knowledge and the ability to retrieve word meaning as `passive/receptive' knowledge. Another distinction is that there is "a difference in knowledge between those who can recall the form or the meaning of a word and those who cannot do this, but can recognize the form or the meaning in a set of options" (p. 206). Accordingly, I tested word retention for four different knowledge 'modalities', (1) active form recall, (2) passive form recall, (3) active meaning recognition, and (4) passive meaning recognition. Although they have an implicational relationship, in the CATSS test the four levels are treated as independent measures (Eckerth & Tavakoli, 2012).

Even though Laufer et al.'s "active/passive" terminology may be misguiding, it is used here to specify the different facets of word knowledge assessed in the test.

The test items for the original test were taken from different levels of the Academic Word List (AWL; Coxhead, 2000). However, in my study the CATSS test needed to be adapted, and the vocabulary items were not chosen according to AWL levels, but according to the selection process described in section 3.5.2. Furthermore, considering Read's (2000) three dimensions of vocabulary testing, parts of my set-up fall into the 'discrete, selective, and context-independent' category (Read, 2000; Read & Chapelle, 2001). Read was critical towards such test types. He objected to the fact that they do not measure beyond the decontextualized items. According to him, considering context does not only make sense because words occur naturally in a context, but it should also generate a positive washback on teaching and learning. However, context-dependent testing is not without criticism.

It is, for instance, not always clear, whether word knowledge or the ability to deduce a word's meaning from context is tested. It is difficult to formulate natural sentences so that a word can be elicited without the task turning into an inferencing task. Naturally, discrete tests measure types of knowledge that are different from context-dependent tests. Being able to

[32] Laufer et al. (2004) deliberately distinguished between 'strength' and 'depth' of knowledge. The CATSS test measures strength of meaning knowledge rather than depth, as depth would include knowledge about aspect such as grammatical features linked to the word or pronunciation, which are not assessed in the test.

recall a word in isolated form is different from being able to deduce the meaning of a word from context. In addition, 'embedded', 'comprehensive', context-dependent' tests would have required a higher level of word knowledge than what is assumed to be known after encountering a word from reading it once (Laufer et al., 2004). An adapted version of the CATSS test therefore was a good compromise, as it combines "measures of different dimensions of word knowledge" (p. 204), and the retrospective interviews add an extra contextual element to the testing set-up, by exploring the background of vocabulary acquisition.

The advantage in Laufer et al.'s (2004) computer-based study was that the computer software omitted those items that had already been successfully known at the highest possible modality, so that they were not tested again. Thus, the students were tested according to their individual proficiency. As it is a diagnostic test, the focus was on the scores of the individual test subjects and showed exactly how well the different items were known. Unfortunately, this scoring system was not possible in my study (see section 3.11.1).

Table 6 illustrates the structure of the CATSS test:

Table 6. CATSS-Test: Tested Types of Vocabulary Knowledge (Laufer et al., 2004:206)

	Recall	**Recognition**
Active **(retrieval of form)**	active recall: supply L2 word form	active recognition: select L2 word form
Passive **(retrieval of meaning)**	passive recall: supply L2 word meaning	passive recognition: select L2 word meaning

Active recall was tested first, followed by passive recall, then active recognition, and finally passive recognition. The particular order of the different word knowledge types moved from the assumed most difficult task (active recall) to the easiest type (passive recognition). Thus, the participants would move along the knowledge continuum in a top-down approach.

3.7.2.2. Pre-test II: Approach

Pre-test II comprised of five parts. The first part, 'original sentence writing' was added to the four parts of the original CATSS-test in an attempt to have one components in which the target words would be handled within a 'natural context'. The participants were provided with a list of the target words and were asked to use the words in sentences. The instructions were the following:

Figure 3. Vocabulary Post-Test Part A – Original Sentence Writing

Matching Exercise A

The following words are all from the text that you have just read. Please try to use them in a full sentence.

Example

flabbergasted (adj.)

When I heard how much money we'd made, I was flabbergasted.

1. ravenous (adj.) ……
2. …

It was expected that this task would give good insights into word knowledge depth. The other four parts of the VPT followed the CATSS-test pattern. All the examples sentences were taken from Laufer et al. (2004:206-08), but the wording was adapted to suit the competence level of the participants.

Part B, the 'active recall' task assessed the ability to retrieve a word form actively. As suggested in Laufer et al. (2004), the first letter of the target word was provided to prevent the participants from supplying non-target words with the same meaning:

Figure 4. Vocabulary Post-Test Part B – Active Form Recall

Matching Exercise B

The following words are all from the text that you have just read. Read the sentences and fill in the blanks.

Example: *Turn into water*: m_____
Correct answer: melt

1. If you have not eaten for a very long time you feel r_____.
2. …

The next part of the test (C) measured the participants' ability to recall the target word form passively. As in the original CATSS test, test-takers had to finish an incomplete sentence in which the target word was embedded. The participants were informed that several answers were possible.

Figure 5. Vocabulary Post-Test Part C – Passive Meaning Recall

Matching Exercise C

Please fill in all the blanks

Example: When something *melts* it turns into _____
Correct possible answers: When something *melts* it turns into water
When something *melts* it turns into liquid

(several correct answers or ways of expression are possible)

1. When you feel *ravenous*, you are _____.
2. …

In part D, students had to choose the word which best matched the definition from five options. The distracter items were semantically unrelated, but chosen from the same frequency level as the correctly matching word. As in Laufer et al. (2004:207). The instructions were the following:

Figure 6. Vocabulary Post-Test Part D – Active Form Recognition

Matching Exercise D

Please underline ONE option

Example:
Turn into water: a. elect, b. blame, c. melt, d. threaten, e. "I don't know"
Correct answer: a. elect, b. blame, c. melt, d. threaten, e. "I don't know"

1. To feel extremely hungry or starving.
a. convenient, b. ravenous c. recreated, d. deserved, e. "I don't know"
2. …

This format tested the ability to actively recognize the target word form. Even though Laufer et al. (2004) found that the two recognition categories (here part D and E) were not significantly different from each other, I found the distinction useful as I was measuring the initial stages of word learning. I was interested in finding out whether different levels of recognition ability could be measured in my data.

Finally, in part E, the correct meaning of the target word had to be chosen from five options, which were all taken from the same frequency level:

Figure 7. Vocabulary Post-Test Part E – Passive Meaning Recognition

Matching Exercise E

Please underline ONE option

Example:

Melt a. choose, b. accuse, c. make threats, d. turn into water, e. "I don't know"
Correct answer: a. choose, b. accuse, c. make threats, d. turn into water, e. "I don't know"

1. *ravenous*
a. marital, b. embarrassing, c. starving d. gradual, e. "I don't know"
2. …

Various problems with this test were anticipated and addressed whenever possible. One concerned the test's complexity. Waring & Nation (2004) found that "using several tests presents a fuller picture of learning and in future work of this kind it will be important to collect data from more than one type of test so that we can better understand what is going on when learners read" (p. 16). However, it also has to be considered that multi-perspective testing with different stages is particularly prone to cause test fatigue and a test training effect. These effects were considered in test design and data analysis. Embedding the texts in classroom teaching was hoped to diminish the training effect. It was also assumed that the text topics were interesting and relevant and that this interest would outweigh test fatigue. The number of target words was kept at a reasonable level so that the willingness to participate in the test would persist. In all five parts of the vocabulary post-test the target words appear in random order to avoid a practice effect. While test fatigue and training effect cannot be fully eradicated, these factors were explored in the retrospective interviews, so that their effect on the vocabulary post-test results could be assessed.

As mentioned above, the active/productive vs. passive/receptive terminology applied by Laufer et al. (2004) was not always fully comprehensible. However, as the active/productive vs. passive/receptive distinction generally is a somewhat vague categorisation (see section 2.2.3), I chose to keep the terms, but to be careful when interpreting the results of the test with regards to what exactly was tested in the different parts of the test.

Another point of criticism concerns the fact that the CATSS set-up includes multiple-choice tests (MCTs). There are several advantages to MCTs and some of the best-known vocabulary tests are MCTs (e.g. the Vocabulary Size Test, Nation & Beglar, 2007 or the Vocabulary Levels Test, Nation 1990; Schmitt et al., 2001). They are easy to administer, and in vocabulary testing in particular, it is possible to test many words at a short time. MCTs are well suited to assess how well learners remember a word form and whether they can distinguish between

word forms (Simard, 2009, Wesche & Paribakth, 1996). However, there are also shortcomings. One is that test items appear decontextualized. Moreover, MCTs work through a process of elimination, which presupposes some knowledge of the distractor words, and which always leaves a certain percentage of chance to simply guessing correctly (Gyllstad et al., 2015; Roediger & Marsh, 2005; Stewart & White, 2011; Stewart, 2014). This can inflate scores considerably.

To diminish these effects, several non-target words from the texts were used as distractor words in the choice words in the VPT. Moreover, the MCTs were only one part of the test. The other VPT parts supplemented their findings and influenced how participants perceived the testing procedure. In the analysis, I took care not to overgeneralise the results to the overall vocabulary knowledge, but remained aware of the fact that these parts of my test merely measured recognition skills. However, in hindsight, I think that choosing different types of vocabulary tests could have made the effect of TIEs on incremental vocabulary acquisition clearer. I would now integrate a more contextualized type of measurement instrument into my testing set-up. For instance, this could be a more systematic questioning of the interviewees' on their specific word knowledge and/or the use of think-aloud protocols while reading.

3.7.2.3. Pre-Test II: Procedure

To retrieve more information about target word knowledge within context and to find the most suitable texts, a second pre-test (part II) was conducted in three secondary-school classes in Copenhagen (N = 62). The learners were in terms of age and level of proficiency similar to the main-study participants. The classes were randomly subdivided into groups of about 10 students. Each of these groups was given one of the chosen texts to read. The texts did not include any textual enhancements.

In the task instructions, the participants were asked to read their text and to highlight 'unknown'[33] words with a coloured pen:

Figure 8. Pre-Test II – Instructions Task 1

1. ***Please read the following text carefully. While reading, please underline or highlight all the words that you do not know or do not understand.***
2. ***After reading, please complete the exercises you will be given. Thank you!***

After they had returned the texts after reading, the learners were given a list with the potential

[33] Even though what one perceives as 'unknown' may differ individually, there was no information added to further clarify this category. This was done to avoid confusion and because the first part of the pre-test provided enough information of this kind.

target words and were asked to give a Danish translation or to formulate a full sentence with each of these words:

Figure 9. Pre-Test II - Instructions Task 2, Example Taken from 'Hollywood' Text

VOCABULARY EXERCISE: 'The Ecology of Hollywood'

Please write a full sentence and/or give a Danish translation of as many of the following words as possible:

1. seething ……
2. entitlement ……

Lastly, the participants had to indicate how 'interesting' they had found reading their text on a scale from 1 to 5 (see section 3.5.1). This provided me with another instrument to choose the most suitable three texts, which turned out to be Text 1 (Horror), Text 2 (Divide), and Text 3 (Equality).

3.7.2.4. Pre-Test Part II: Analysis And Results

While the second part of the pre-test confirmed many of the potential target words, other words which had been assumed to be unfamiliar to the students were not marked as unknown. Subsequently, these words were not used as target words.

To distinguish between correct and incorrect translations/definitions of the target words, book and online dictionaries (e.g. Longman Dictionary of Contemporary English) were consulted, as well as the researcher's own knowledge of English and Danish, and Danish/English bilingual informants, who included some of the teachers.

The following examples illustrate the spectrum of the responses: While 'a place' was counted as unknown/incorrect as it was too vague for the target word *realm*, 'something with resident/home' was marked as known/correct for *residential*, because it illustrates that the learner was aware of the specific semantic field/meaning conveyed in the word. One particularly striking aspect was that learners often seem to read texts while following misconceptions of the meaning of many words. This became obvious as low-frequency words like *disparate* were only marked as 'unknown' by relatively few students, but almost none of them managed to provide a correct translation / definition. They probably mistakenly assumed they knew the word, but mixed it up with 'desperate'[34].

[34] Similarly *precept*, often translated as 'perception', *trough* vs. through, *cease* vs. seize, *divert* vs. divide.

3.7.3. Delayed Vocabulary Post-Test

Research on language learning is also concerned with how well the learned items are retained over time. This is especially true if conclusions are drawn for the effectiveness of task design for L2 classrooms. However, in many studies the long-term effects of interventions are not examined. In my study, a delayed post-test was conducted four weeks after the last treatment. The delayed post-test comprised of a similar test set-up as used in the vocabulary post-test, except for the fact that the target words were not split up between the texts.

As already mentioned, the VPT were only one part of the post-reading treatment. In order to get a more holistic insight into the vocabulary acquisition processes and to understand the effects of the TIEs better, interviews were conducted. The goals and the structure of these interviews are explained in the following sections.

3.8. Retrospective Interviews

In early TIE research it was claimed that “input enhancement implies that we can manipulate aspects of the input but make no further assumptions about the consequences of that input on the learner” (Sharwood Smith, 1993:176). In my study, I decided to add retrospective interviews as an attempt to reveal something about these consequences. I wanted to explore how exactly individual learners approach enhanced reading materials and how they tailor tasks to their needs. I was also interested in how well the target vocabulary had been acquired and the respondents’ attitude towards testing and to working with TIEs. While these aspects might seem distinct, in the interview process, they often overlapped.

As mentioned above, I partly follow the methodological design of previous research in the field to make is easier to compare findings. However, replicating existing methodology does not come without problems. As Han et al. (2008) point out, there are methodological issues regarding how TIEs have been investigated. They criticise that researchers have more or less equated the effectiveness of TIEs with their ability to generate acquisition, where acquisition is associated mainly with improved accuracy in production. Moreover, they critique that in most previous studies “statistical significance served as a prime, if not the only, indicator of whether or not textual enhancement was effective” (Han et al., 2008:599). Several recent vocabulary and TIE-related studies took a more holistic approach and included qualitative data (Cheng & Good, 2009; Gettys et al., 2001; Ko, 2012; O’Donnell, 2012). However, they almost exclusively worked with questionnaire data. While there certainly are advantages to survey data, there

are also limitations. I try to address some of these issues in my study and hope to provide insights that go beyond what has been achieved in most of the previous studies. Including retrospective interviews is a trend particularly in vocabulary acquisition studies. For instance, Pellicer-Sánchez and Schmitt (2010) used such interviews to explore how well the target vocabulary was known. They asked their participants to say everything they knew about the meaning of the target words and also examined test attitude.

3.8.1. Aims

As explained above, I decided to include interviews in the research design to complement the quantitative data collected in the VPT. These interviews aimed at examining four aspects in particular, the interviewees' task approach, effects of the reading conditions, their word knowledge, and attitudes.

Task Approach

How a task is approached by the individual learner is decisive for successful task completion (Eckerth, 2009), because "though the teacher [or researcher] may control the experiences the learner is exposed to, it is the learner who selects what is learnt from them" (Dakin, 1973:16). Considering that "learners are capable of playing havoc with even the most carefully designed and much used task" (Breen, 1987:23 cited in Block, 1996:168), it is important to ask whether, for what and how TIEs are used in a task. A better understanding of how learners exploit tasks beyond their original focus and how personal learning dispositions contribute to task approach can help designing TIE tasks for vocabulary acquisition. My interviews aim to investigate this. The perception of what is important and relevant to focus on in a task may vary widely between learners. The interview questions take this into account, for instance by exploring how far 'incidental' learning is actually 'incidental'. Responses may thus help exploring the difference between 'incidental' and 'intentional' word learning. The interview questions will evolve around what the learners struggling with in the task, what they did first or last, how they worked out unknown words; what was going through their mind, what stuck in their mind and what was ignored[35].

[35] For the 'interview guide' with possible questions, please see Figure Appendix 10.

TIEs - The Reading Conditions

In the interviews, I was also interested in gaining insights into whether and how learners use the different TIE-types, therefore, questions were asked to reveal their way of working with them. However, this is a particularly difficult issue to measure. Danish learners are so much used to working with TIEs that many of them use different enhancement-forms automatically. Therefore, interviewing them about this means measuring levels of awareness. The intention was to find out in how far they consciously work with the different text forms and through indirect questioning bring less conscious approaches to light. This was done through subtle questions such as 'Was there anything in the text that helped you understand it better?' rather than directly inquiring whether and how they had used the glosses. This is related to the problematic 'incidental' versus 'intentional' distinction, as it must be assumed that consciously working with target words by using textual enhancements when completing a reading task is different from only using them incidentally.

Word Knowledge

In the interviews I explored word knowledge beyond what was revealed in the VPT. By probing, for example which facets of vocabulary knowledge had been acquired (e.g. recognition versus productive skills), it was hoped that the interview data might give an insight into the incremental nature of vocabulary knowledge.

The word-knowledge related questions were based on Nation's (2001) list of questions about what is involved in knowing a word (p. 27). During the interviews, the post-test results were available, so that the participants could comment on their performance and give examples if they wished. The known/unknown target words were used as prompts and the respondents were asked to state related word-associations or to comment on why or how they remembered the target word. To investigate the depth of understanding, the participants were asked how the word could be used in a sentence or which collocations or associations were related to it.

Learner Attitudes

The next part of the interviews probed the participants' attitudes regarding the use of TIEs and towards being tested.

This was expected to give further insights into how TIEs are used in practice and for which purposes. Danish students are used to working with enhanced texts, so it was assumed that

many of them pursued specific, potentially individualized routines. Identifying these might explain some of the quantitative findings of this study.

I was also interested in the students' attitude to testing, which particularly in a school context could play an important role for findings: Dörnyei (2007:189) assumed that "It is highly unlikely that every student will do his/her best for a project in which they have little interest and which has no direct bearing on their school grades" (similar Mackey & Gass, 2005). How seriously participants take a study and how they feel about being tested has been found to have an impact on the outcome of a study and therefore should be taken into account in research (Dörnyei & Kormos, 2000; Mackey & Gass, 2005). Mackey and Gass (2005) recommended to gather post-experiment data as helpful for interpreting results more holistically and pointed out that interviews in particular can serve as an in-depth basis for understanding learning processes.

It has been argued that interviews should elicit learner perspectives and voices (Block, 1996; Strike, 2006). I did this in my interviews. Additionally, this meant triangulating the test results and therefore enhancing the overall quality of the research.

3.8.2. The Form Of The Interviews

Ideally, all interviews were to cover the four areas of interest in my study (task approach, approach to the use of TIEs, word knowledge, test attitudes). Therefore, the most suitable form was the 'semi-structured interview'. This slightly vague term covers a wide range of approaches. In my study it describes interviews in which the researcher uses a written list of questions as aide-mémoire, while still having the freedom to digress and probe for more information (Mackey, 2005) and where the style of questioning is informal (Bryman, 2008). Thus, the researcher knows what topics need to be covered and to a large extent what questions need to be asked (Richards, 2009; Brenner, 2006). However, the questions can vary in sequence, and wording, and further questions can be added (Bryman, 2008). Another reason why the semi-structured form, as opposed to for instance an ethnographic form of interviewing, was chosen for this study was the research context: This is a classroom study, trying to explore learning in a natural environment without interfering too much with the normal learning situation. Therefore, more artificial and disruptive methods, like think-aloud protocols, were not regarded as suitable. Furthermore, the interviews had to be kept as short as possible, so as not to keep the students away from their lesson for too long. Moreover, a more structured form of interview is likely to be more successful at keeping interviewer variability at bay and therefore better suited than an ethnographic form of interview (Bryman, 2008). However, in a semi-structured interview, the interviewer still needs to be flexible and to "probe some aspects in depth and, where necessary, to let the respondent lead in much the same way as in an open interview" (Richards, 2009:186). The main focus must be on anticipating how

interviewees understand the issues under discussion and what they may perceive as important (Bryman, 2008). When phrasing the questions, the goal was not to use a standardised interview form, i.e. not asking only closed questions, but being open for individual responses while still making the different responses comparable. I tried to use open-ended questions, of the kind suggested by Brenner (2006:362-3): "*how* and *what* question that cue the informants to give their perspective in their own words". At the same time, I took trouble to ensure that the questions were consistent from one respondent to another and as free of bias as possible so that answers from a group of respondents could be codified in such a way as to be analysed effectively (Keats, 2000). Striking this balance was one of the challenges. As a help, a short 'interview guide' was designed, which divided the interview into topics, contained initial wording of questions, and a list of areas to be explored with each informant (Figure Appendix 10; see Brenner, 2006; Richards, 2009; Bryman, 2008).

Interview participation was voluntary. However, respondent self-selection may be problematic. It may affect validity, because the resulting sample may not be similar to the target population. Dörnyei (2007:100) pointed out that "volunteers may be different from non-volunteers in their aptitude, motivation or some other basic characteristics". To avoid this situation, I explicitly invited less confident students to participate. Lastly, some interviewees were not very confident speakers of English and used the interviews, for example, to talk about general learning difficulties.

In order to establish a good rapport with the test subjects I made an effort to create a friendly atmosphere and gave all participants the chance to ask questions. However, as the purpose of the interview was to get a realistic impression of the participants' learning, to not "change the respondent's attitudes and behaviour but to reveal it", professional distance had to be kept (Keats, 2000:7). Keeping this distance was one of the challenges of the interviews.

The precise course of the interviews varied, depending on factors like the confidence of the interviewee or the location. I briefly explained the purpose of the study again and stressed the ethical framework (anonymity, handling of recordings). With less confident students I repeated the offer that using their L2 (English) or their L1 (Danish) or a mix of both would be acceptable. Then, the conversation continued with some small talk to make the interviewee feel comfortable. When all practical problems were sorted out[36] the interview was opened with a 'grand tour question', which is typically an opening question that asks the informant to give a broad description of a particular topic (Brenner, 2006). While it was, as mentioned above, important to cover the four areas of interests in the procedure, the intention was still to let the

[36] For example finding the interviewees written tests.

respondents give direction to the interview process. Therefore, open-ended and closed questions were mixed; the sequence of the questions and topics and the length of the responses varied enormously. All interviews were conducted directly following the reading/testing session and lasted for approximately 15 minutes.

Throughout the interviews notes were taken in addition to the sound recordings. These were about any noteworthy issues, like the location of the interview, the wording of a question or any issues occurring on a specific day that may have affected the interview. For instance, this was the case when one of the classes returned from a PE lesson and was particularly exhausted. Sometimes these notes led to an adaptation of the interview guide. Other times specific questions were not clear for the interviewees and had to be rephrased or an interesting issue emerged that had not been anticipated in the planning. In the data analysis, however, I found that I should have adapted the interview guide more thoroughly in order to avoid some of the shortcomings in the data (e.g. vague responses, which were not clarified). However, as the interviews were conducted in a rather short period, there was seldom and opportunity to consider the quality of the already collected data or to listen to the interviews carefully before going ahead with the next ones.

3.9. Pilot Study

To ensure the feasibility and validity of the design, the main study was preceded by a pilot study in which the reading tasks, vocabulary post-tests and interviews were tested. The following procedure was used:

- General information about the overall goal of the study was provided and questions clarified. Volunteers were invited to participate in post-treatment interviews, which would be used to collect general information about their task approach. (10 mins)
- Participants were handed a set of papers (cover with the instructions, comprehension questions, the text and the post-reading activities). They were asked to read the instructions in order to clarify questions. The post-test was not mentioned. (5 min)
- The participants worked through the reading tasks (45-60 min) and returned the papers when done.
- The participants were given the various parts of the vocabulary test, one sheet at a time, and returned each sheet when done. (15 mins)
- The interviews were conducted with volunteers. (15-20 mins)

As in the pre-tests, to avoid a practise effect, the pilot study was conducted with three classes (N = 55) in a school with students of the same age group and competence level as the students participating in the main study. The results of the pilot study were not included in the data corpus of the main study.

In the following section, I report the findings of the pilot study with respect to the test content, its procedure and regarding the interviews. Changes to the original set-up are explained.

3.9.1. Test Content

The data collected in the pilot study and the notes taken while conducting the tests showed that the methodological design needed to undergo several changes. These changes concerned the reading tasks and vocabulary post-test as well as the overall procedure of the test treatment.

The students' reactions to the reading tasks and also the interviews revealed that the multiple-choice reading comprehension questions in the reading tasks were not always clear enough to be answered correctly. Some students chose to tick several options. None of these participants regarded this as 'problematic'. One of the two interviewees pointed out, that he had found it beneficial to have several answer options as they enforced thinking about the answers. Accordingly, I adapted the wording of several comprehension questions and post-reading activities.

Several changes concerned the chosen target words and the set-up of the vocabulary post-test. The distractors for the two passive recognition modalities had been chosen from the same (or higher) frequency level as the stimulus word. It became clear that they were likely to vary in terms of their plausibility as synonyms for the target word, thus making items more or less difficult than expected. This problem is commonly encountered when designing MCTs (Laufer et al., 2004). I therefore substituted several implausible distractor words with more suitable items and tried to find distractors within the reading texts in order to reduce guessing.

I also found that generally, to avoid confusion, the target words in one text should not start with the same letter (as for example *indulge* and *impose* in the Equality-text). The texts were adapted accordingly. Similarly, the target words *notion* and *perception* (Horror-text) were too similar in meaning and *perception* was therefore substituted by a different word. In addition, the options for the 'solution' words were found to be too alike (dangerous versus evil), and therefore had to be substituted by the target word *malevolent* (Horror-text).

Furthermore, to avoid 'test' terminology and to make the tasks more accessible for students and teachers, in the instructions the vocabulary post-tests were called 'matching exercises'.

Unlike the CATSS vocabulary test, for the pilot study 'original sentence writing' (part A) had been added as a fifth part to the VPT. However, Part A and B of the VPT presented various problems. In these parts of the VPT, participants were A) asked to use the target words to write original sentences, and B) to complete incomplete sentences when prompted with the first letter of the target word. Both of these tasks presented similar problems. Part A was particularly challenging and time-consuming for the participants and therefore de-motivating. In addition, the scoring system and analysis were ambiguous. It was nearly impossible to clearly define what a 'correct' original sentence is which at the same time reveals true knowledge of a particular word. The data yielded too little insight and the scores were extremely low. Therefore, they were not investigated further. Concerning part B, Laufer et al. (2004) had claimed that providing the first letter of a target word would prevent test subjects from using non-target words that have the same meaning. This, however, was not the case in my data. Participants produced grammatically and semantically correct sentences, but did not necessarily use the target words. Instead, they filled the gaps with whatever word they thought fit best (e.g. 'Her husband was gone, so she had to *rear* the boy all by herself.' versus 'Her husband was gone, so she had to *raise* the boy all by herself.') which made fishing for a particular target word difficult and artificial. Secondly, the results were overall very meagre, many participants did not enter any information in this part of the test. I assume that it was too challenging.

After careful consideration, I decided to exclude part A and B from the VPT set-up and not to test sentence writing and active recall in the main data collection. This may seem unfortunate with respect to striking a balance between context-independent and context-dependent types of vocabulary measures (Read & Chapelle, 2001). It further created another problem. Nation (2001:30) recommends that when comparing receptive and productive learning modes, the two test items should either be both recognition items or both recall items. If both types are mixed, then it is "impossible to tell how much the difference in scores is a result of the productive/receptive distinction or the recognition/recall distinction". By eliminating the productive type of form recall in my study, the tested modes are out of balance for the recall mode. It was therefore important to treat the 'passive meaning recall' part of the test especially carefully and not to overinterpret the findings linked to this part of the test.

However, the adapted form of the CATSS test, i.e. a test that still includes different dimensions of word knowledge was considered a good compromise between fully contextualized and decontextualized one-dimensional tests (e.g., Laufer & Nation, 1995; Meara & Buxton, 1987). Eliminating part A and B also had the advantage that the post-test was shortened. The results of the pilot study showed that these productive parts of the needed to be replaced by another measure of recognition. The chosen one is based on a study by Henriksen and Haastrup (1998) in which they explore the continuum leading from partial to

precise word understanding. Their test is based on the Vocabulary Knowledge Scale (Paribakht & Wesche, 1996):

Figure 10. Vocabulary Post-Test Part A - Adapted Test Format

Exercise A

Do you recognize any of these words? Please tick the correct box.

	word	I know the word	the word looks familiar, but I do not remember what it means	I don't know the word
1	profound			
2	tumpy			
…	…			

This self-reporting test measures the learners' sensitivity towards word recognition and gives them the chance to describe what they mean by 'recognize'. In the task, they are asked to distinguish between real English words from English sounding/looking non-words such as *tumpy*. Thus, informants are not encouraged to reflect on meaning, but rather to rely on formal word features such as orthography. Ellis (1994) regards vocabulary acquisition regarding the learning of such surface forms of language as an implicitly acquired skill, whereas mapping meaning onto form as a cognitive skill that is acquired explicitly. Haastrup's and Henriksen's (1998) test is more sophisticated than other formats, for instance yes/no tests, as it has the potential to assess different levels of partial knowledge. In their study, similar to mine, the goal was to "trace the initial phase of learning" as "formal features are interwoven with meaning aspects and productive skills with receptive skills" (Haastrup & Henriksen, 1998:102). In other words, they hypothesized that partial knowledge in the form of recognition was a necessary prerequisite for precise knowledge and found this confirmed in their results. Tracing this assumption was helpful also for exploring the questions under investigation in my project.

In my test, the 'I don't know' option was added in order to decrease guessing effects and to increase test reliability (Pellicer-Sanchez & Schmitt, 2010; Stewart, 2014; Zhang, 2013).

Piloting also showed that the instructions to part C were confusing for some participants. As the given example dealt with changing of aggregate states ("When something melts it turns into …"), some assumed that they were required to find similar examples related to physical change for the given words. Others thought that they had to provide several answers and felt unable to cope. On top of that, one student was not familiar with the word 'to paraphrase' and

therefore did not know what he was expected to do. Therefore, the instructions for the main data collection were simplified, as shown in Figure 12:

Figure 11. Vocabulary Post-Test Part C – Passive Meaning Recall: Adapted Instructions

Please find an equivalent or similar words or rephrase the listed words (see the example below). Several correct answers may be possible.

1. entail (verb) - ……
2. besetting (adj.) -……

This was a clearer way of asking for the same skill: filling in the blank required the participants to provide a conceptual description of the target word by means of a synonym or semantically related words. Even though it was unfortunate that the target words were provided here in a decontextualized form, these adaptations were necessary in order to match the task to the participants' proficiency level.

Some options in the recognition parts of the test offered as 'definitions' or 'synonyms', covered only the specific meaning the word has in the particular text, for example *inflexible* for *rigid* or *approve* for *indulge*. This might be misleading for learners as they might have encountered the word in a different context and might find the given option different from their previous understanding. However, this problem reflects natural word learning situations. We always learn the meaning of new words in specific contexts. This meaning will very likely only cover a specific, restricted meaning of this word and will not be a general 'dictionary' understanding of this word.

The pilot study showed that many students did not use the given information about which part of speech a target word belonged to, i.e.

to tailor (v)

encounter (n).

Instead, they used the words in whatever form they deemed fit in the vocabulary post-test. Nevertheless, this information was kept in the test set-up, as it might be helpful for some students.

The vocabulary post-test scores largely confirmed results of previous related studies. Generally, the enhanced words yielded higher scores than the unenhanced target words.

The final vocabulary test set-up was as follows:

1. Part A – passive form recognition (indicating the familiarity of word form and meaning)
2. Part B - active form recognition (selecting the target word form that matches with a given definition from a choice of four)
3. Part C - passive meaning recall (retreiving a word meaning equivalent or similar to the provided target word form)
4. Part D - passive meaning recognition (selecting a matching (near-) synonym to the given target word form)

The order chosen for these different parts of the test deviated from the Laufer et al. (2004) study. While they investigated the sequence of the test levels, I had to consider slightly different concerns, namely the focus on word form and meaning. Part A of my test set-up did not provide any clues beyond word form. In part B, clues to word form and meaning were given. If the testee had no knowledge of the word's meaning, therefore, this part could not be completed successfully. In part C again only word form was provided but meaning had to be recalled this time. After having encountered a word only once in a text, the probability of being able to recall word meaning is slim. Therefore, in my study, it seemed most important to get a differentiated view of the lower range of vocabulary knowledge assessed in the recognition parts of the test, while at the same time the participants were not to be over-sensitised to the target words to avoid a training effect. Therefore, passive recall was placed before the final stage of the test, passive recognition (part D).

3.9.2. Test Procedure

The pilot study showed that all task instructions and explanations concerning the study had to be introduced before the first test treatment in a short extra session. This was necessary to ensure sufficient time for all parts of the procedure. It was important not to conduct this introduction too long before the testing itself, as the students might forget important details. Participants then commenced the task immediately. Unlike Laufer et al. (2004), my participants were only given one part of the VPT at a time to work on. Thus, they could focus on each separate assignment and were not tempted to jump between the lists to look for clues, which would have affected results.

Students, who finished earlier presented a problem. While the others were still working on their reading and follow-up tasks, these students were ready to work on the vocabulary post-test and would have revealed the nature of the VPT and details of the target words: The early finishers could easily peek into the text of the reader sitting next to them and see a target word's definition there. Therefore, those who finished early were instead asked to check their responses to the pre- and post-reading exercises again.

Specific challenges arose regarding the reading tasks and the VPT: In a study investigating the optimal vocabulary knowledge coverage for successful comprehension, Hu and Nation (2000) took the texts away from their participants for the comprehension part of their tests. In my study, however, I was not measuring reading comprehension as such and explicitly wanted the students to engage with the text naturally while trying to answer the reading comprehension questions. Therefore, they were allowed to keep the texts while answering the comprehension questions and completing the post-reading activities. However, before starting the VPT, they had to hand in all the papers.

Even though my participants had been informed about how many parts there were to the test treatment, some found it difficult to manage time when not knowing exactly what to expect. Even though they were advised not to get stuck and instead to move on to the next, potentially easier, task, most of them tried to complete all the parts successfully. The students' ambition to perform well seemed to potentially interfere with the time restrictions of the test procedure. Therefore I decided that in the main study the teacher would periodically remind the students of the remaining time (half hour, 20, 10, 2 minutes).

3.9.3. Retrospective Interviews

The two interviews conducted in the pilot study revealed that some students were only vaguely aware of their word learning behaviour. To ensure that the interviewees could reflect on the details of their thoughts and actions, I decided to conduct the interviews as soon possible after the test completion. It was further important to explain and stick to the overall structure of the interview guide so that the interviewees would not be confused. At the same time I had to be open for comments that may break the structure planned in the interview guide.

The interview responses revealed that task completion really was a highly individualized process and that task design could not necessarily predict task approach. Responses to questions concerning vocabulary knowledge showed a rich variety of used strategies. Overall, the interviews promised interesting findings for the questions pursued in this study, confirming the suitability of interviews for exploring the research questions in my study.

3.10. Main Data Collection

Following the analysis of the data from both pre-tests and the pilot study, the main data collection was conducted over a three months period, between October and December 2010, with the delayed post-tests following in January 2011. However, not all participants were present in all three test sessions, so the actual number of test subjects is slightly lower. This

number includes 16 interviewees, whose scores were removed from the main data set, in order not to distort the test data[37]. The test sessions were conducted once a week, in three consecutive weeks. Data was collected in 12 EFL-classes, two of which were control groups. Due to time constraints, two of the 12 tested classes participated in only two test sessions.

3.10.1. The Reading Tasks And Vocabulary Post-Test

Table 6 shows how the numbers of participants are distributed with regards to the different reading conditions and in how many classes the tests were conducted (2-3). Cell size for each condition with each text varies between N=41 and N=71. Due to practical circumstances, this rather large variation in numbers could not be avoided.

Table 7. Number of Participants Completing the Reading Tasks

'Horror' text	N	# of classes tested	'Divide' text	N	# of classes tested	'Equality' text	N	# of classes tested
Condition 1 (bold-printing)	41	2	Condition 1	65	2	Condition 1	71	3
Condition 2 (glossing)	45	2	Condition 2	67	3	Condition 2	48	2
Condition 3 (bold-printing + glossing)	66	3	Condition 3	68	3	Condition 3	56	3
T1[38] Control (unenhanced text)	51	3	T2 Control	47	2	T3 Control	48	2

All teachers were present during at least part of the reading task and testing process. However, they stayed in the background and were not involved beyond helping to handle the papers.

The time spent on completing the tasks varied slightly from class to class and considerably from session to session. Naturally, the first testing session, which also included an introduction to the purpose and an explanation of the ethical considerations in the study, took the longest time (appr. 80 min), as students had to familiarize themselves with the task format.

3.10.2. The Interviews

Interview data was collected from one or two volunteers from each class after each reading/testing session. Finding these volunteers was not a problem. Usually, 1-3 students

[37] Contrary to the other participants, and because they respond to very specific questions about the vocabulary in the texts, the interviewees are unduly aware of the focus of this study. Therefore their scores may be very different from those of the other test subjects.
[38] T = Text, e.g. T1 = Text 1

were interested in talking about their reading. When there were more volunteers than needed, the students came to an agreement about who should be first. Usually, the problem solved itself when students dropped out or forgot to come along.

Overall there were 16 interviewees, and 39 interviews were conducted. Not all of them were present for all three interview sessions. Limitations of the interview data and the interviewing technique are discussed below (see section 7.3).

3.11. Data Analysis

As the data collected in my study was so diverse, a variety of analytical tools had to be used. The following sections explain which types of analysis were chosen.

3.11.1. Vocabulary Post-test Analysis

I used statistical analysis to find answers to some of the aspects investigated in the research questions. I was interested in whether the TIEs had an effect on the outcomes of the VPT and whether these effects differed regarding the different types of word knowledge. Statistical analysis can also shed light on other factors that influence vocabulary learning. Accordingly, the analysis had to yield results that would allow the comparison of different TIE types ('groups') and show whether and how the different variables were related, interdependent and/or interacted.

Preparing the vocabulary test data for analysis required checking, organizing, cleaning and (numerical) coding, i.e. defining each variable and then compiling coding specifications for every possible 'value' that the particular variable could take (Dörnyei, 2007; Phakiti, 2010). In this study, the vocabulary post-test scores, for example, were summed up as '1 full point' for a correctly known word, '0 points' if the word was unknown or incorrect[39]. The values were then statistically examined using SPSS, Stata and the www.laerd.com website.

My study investigates the effects of factors such as TIEs or learning approach ('independent variables') on the initial acquisition of vocabulary from reading ('dependent variable'). Regarding its statistical categorisation, the dependent, or 'outcome' variable, vocabulary acquisition, measured in the form of VPT scores, is not easily categorized: The VPT scores have possible outcomes from 0-10 points, all of which have the same difference between

[39] This scoring system was necessarily a simplification of 'vocabulary. In order to make a quantifiable analysis possible, pre-conceived categories had to be used regarding to what counted as correct or incorrect.

them. Therefore, they are discrete data that can be ranked in meaningful intervals and measured on a ratio scale. For the purpose of the statistical analysis all target words were regarded as being of equal value, because the research set-up could not test for word difficulty. However, unusually for interval data, the scores contained a natural zero point, i.e. scoring '0' meant remembering zero words. Additionally, the scores clearly had ordinal features, there is a monotone relation between categories, so that larger values represent 'better' (Bryman, 2012; Field, 2005). Usually, however, with ordinal data no precise differences between the ranks are assumed, but are rather ordered categories which cannot be precisely measured. At the same time, however, with 11 possible data points (0, 1, 2, … 10), the data has continuous features (Harwell & Gatti, 2001). On the other hand, to treat the data as continuous in nature, the number of choices is too limited. Therefore, to find the most appropriate categorization for this variable, I looked at the overall nature of the data: It seemed most appropriate to treat the VPT scores as discrete ordinal ratio scale data, as the 'continuous' aspects were too limited. Regarding the data as 'ordinal' also makes sense because ranking the scores into 11 data points ordered from the lowest to the highest seems the most natural approach. This reflects how a schoolteacher would treat test outcomes[40]. It was hoped that potential shortcomings of these categorizations would be alleviated by combining the quantitative data with the interviews.

Session, school, gender, text, and language background were included as independent variables. Unfortunately, no access was given to record further possibly details like the participants' socio-economic background, their age or their general performance in school.

Session was included as a variable for three reasons. The first concerns test familiarity and test fatigue. When participants are tested three times, test familiarity and test fatigue might influence the way they work and therefore have an impact on the test outcomes (Eckerth & Tavakoli, 2012). The second reason concerns the difference between incidental and intentional learning. The students' growing awareness of the test procedure is likely to sway how the concept of incidental versus intentional learning has to be investigated. The factor session can indicate incidental versus intentional learning. It is likely that vocabulary learning in the first reading/testing session was incidental, but that the participants increasingly approached the unknown vocabulary intentionally. The third reason to include the variable session was that the data were investigated for each session separately, but also in a pooled data set. Pooling data assumes a within-subjects/repeated measures design. This has the advantage of holding subject variables constant and increasing statistical power by reducing

[40] It does not reflect, however, the way vocabulary acquisition is seen as continuous and incremental. Statistical analysis alone here is limited for comprehensive interpretation.

random variation (Howell, 2013; Maggetti, Radaelli, & Gilardi, 2012). However, working with a within-subject test design can also be problematic. Testing the same participants across factors introduces potential threats to validity through imitation of treatments, maturation and other time sensitive effects (e.g. test fatigue), and through testing effects (Eckerth & Tavakoli, 2012; Howell, 2013; Roediger & Karpicke, 2006). Statistical indication of whether these effects are connected to vocabulary learning seems helpful. This was achieved by using a regression model (see section 4.6.3). These three reasons strongly suggested the inclusion of the variable session in the statistical analysis.

Information about school and gender was also recorded. The participants represented the population in terms of age and proficiency, which was measured in years of EFL instruction in school. Therefore, these two variables were assumed not to have an impact. Still, they were recorded to minimize the error term in the statistical regression model.

Similarly, text was used as a predictor variable in the regression model, but was assumed not to play a differentiating role. Three texts of similar difficulty and structure were chosen together with the teachers. However, equivalent levels of text difficulty could not be guaranteed and some interview responses suggest that Text 1 was somewhat easier than the other two texts (see 4.6.3). Accordingly, the factor text was included to make sure that text difficulty had not skewed the test outcomes.

The participants were also asked about their language background. Twenty-eight students noted that Danish was not their (only) L1, but none of them had English as their L1 or L2. The purpose was to find out whether students with an L1 other than their ambient language performed differently from the others in the VPTs.

Theses variables were examined not only regarding their effect on vocabulary acquisition, but also regarding their potential effect on each other. Their inclusion provides potentially relevant information about the learning process. Thus, they might help answering the third research question of which other factors play a role when vocabulary is learned from reading enhanced texts.

These independent variables are all ‘nominal’, ‘categorical’ variables. In order to include these data into the statistical analysis, numbers were randomly allocated to the different categories (e.g. Text 1, School 2).

As a starting point in the analysis, the vocabulary post-test data were described and the nature of their distribution was examined. These calculations showed that the data were not normally distributed and therefore non-parametric tools had to be used. However, to be able to “say something about possible general lessons that may be drawn” from the data, inferential statistics were also used (Dörnyei, 2007:209).

As outlined in section 3.9, the pilot study had shown that after the first testing session, some participants were increasingly aware of the purpose of the study. This might influence the outcomes of the VPT and had to be considered in the analysis and interpretation. As a way of addressing this problem, instead of analysing the full data set as a unit, it seemed more sensible to split it up into different tiers. This approach is illustrated in Table 8:

Table 8. Analytical Tiers Applied In My Study

tier 1	**test session 1 data**	**→ results?**
tier 2	test session 2 data	→ results?
tier 3	test session 3 data	→ results?
tier 4	data from all three test sessions pooled	→ results?

In the first tier, data collected in the first test session was analysed across classes, texts and reading conditions. Thus, data untainted by a training effect or test awareness was investigated. Similarly, in tier two and three I analysed the data from the second and third testing sessions respectively. In the analysis of the data, it turned out that this step was well justified. The training effect brought about by the repeated testing was observed in the data collected in the second and third testing session. Due to the repetitive structure of the VPT, participants became familiar with the procedure and developed a testing 'routine'. The interviews confirmed this. However, by splitting the data, I could take the potentially increasing training effect into account, it enabled me to answer questions regarding the results when looking at Session 1 data only and could compare this with data collected in Session 2 and 3. This gave insights into the effect of the repeated testing and allowed interpretation regarding incidental and intentional vocabulary learning. In the final tier, the data from the three sessions were pooled to facilitate an inter-individual analysis, in which the scores from all participants for all three conditions are compared.

In order to explore significant group differences between the different treatment groups the non-parametric Kruskal-Wallis H Test was chosen for the first three tiers of analysis. Post-hoc tests such as the Mann-Whitney U test with a Bonferroni had to be conducted to show where the significant differences between the found group differences lay. The Bonferroni correction controls the familywise error through pairwise comparisons that are conducted by correcting the level of significance for each test such that the overall Type 1 error rate (α) across all comparisons remains at .05 (Howell, 2011; Rasinger, 2010; Salkind, 2008).

In the fourth tier, where the data from all sessions were pooled, the set-up had changed to a within-subject design, and the Friedman's Test and Wilcoxon W post-hoc test were chosen

to analyse the VPT scores.

To further examine the relationship between the variables (e.g. session, school), a correlation analysis was applied (Bryman, 2008; Mackey & Gass, 2005). This allowed me to examine different variables and to evaluate the strength and direction of their relationship or association with each other (Dörnyei, 2007, Rasinger, 2008). However, this only applies to linear relationships. To take the complex relationships and dependencies between the different influential variables in my study into account, a multiple correlation model, the Kendall's Tau correlation coefficient, was used (Field, 2005). Such multiple correlations, allow computing the correlation between one variable and a group of variables like for example vocabulary acquisition, input enhancement, and reading session.

Correlation analysis can identify relationships between variables, but not their causes. In my study, causation and direction of effects were identified by regression analysis. Due to the complexity and assumed interdependency of the involved variables, a multiple logistic regression model was used.

A logistic regression models the logit-transformed probability as a linear relationship with the predictor variables. This makes understanding the coefficients in a logistic regression less straightforward than with a simple linear regression. "The logistic regression coefficients give the change in the log-odds of the outcome for a one unit increase in the predictor variable" rather than a one-to-one linear relation between variables ([41]http://www.ats.ucla.edu/stat/stata/dae/logit.htm; see also Field, 2005). It should, however, be remembered that the values do not show causation as such, but rather give an estimation of the predictive power, the strength of this specific variable.

To analyse the data collected in the delayed VPT, which had to take the factor 'time' into account, and to allow for a comparison with the pooled data of the main data collection, a Tobit regression model was chosen. As mentioned above, more details about the choice and nature of the statistical tools are to be found in the Quantitative Findings Chapter and in the Appendix.

3.11.2. Interview Analysis

First, I give a brief overview of the theoretical principles that were considered, then I outline how the interview data were coded and analysed.

Considering the research questions, the major challenge with the interviews seemed to be covering all the issues I was interested in while remaining open for themes outside this framework. Therefore, in order to avoid a pre-structured 'questioning' of students, a custom-

[41] http://www.ats.ucla.edu/stat/stata/dae/logit.htm, accessed on 9/3, 2015.

made mix of a priori and grounded theory approach, a data-led coding system was applied (Brenner, 2006, Chamaz, 2000; Dörnyei, 2007; Mackey & Gass, 2005). 'Data-led' means that the insights were derived inductively, i.e. "obtained gradually from the data" in an attempt to understand informants on their own terms and to explore the data openly (Pope et al., 2000:114; Holliday, 2010). According to Mackey and Gass (2005:241), such 'open' coding has become a common practice: "The schemes for qualitative coding generally emerge from the data rather than being decided on and pre-imposed prior to the data being collected or coded". In this inductive approach, a researcher attempts to describe the categories that emerge from the data during the analytical process (Brenner, 2006). For an effective analysis, qualitative data needs to be carefully prepared by transcribing, describing, coding, and recoding. Due to this character of analysis, rigor is more important here than in other forms of analysis.

The 'data-led-approach' was the principle that shaped my analysis. One feature that most data-driven analytical approaches have in common is that coding should be circular/iterative i.e. a process that is not sufficiently complete after only one examination. This needs to be so to be open to and to consider newly emerging codes. When applying such a coding system, categories are developed through an 'iterative', non-linear approach to data: depending on the emergent results, moving back and forth between data collection, data analysis and data interpretation is necessary (Dörnyei, 2007). In order to come to well-founded conclusions, my interview data was coded by looking for anything pertinent to the research question. Practically, this involved decisions about how to classify or categorize particular pieces or parts of data, through coding according to key words that turned up in the data. This entailed checking how these were distributed in the data, determining themes, constructing an argument, going back to the data, reviewing the codes (Brenner, 2006; Holliday, 2010). Often, in the 'iterative' approach the data collection itself becomes analytical. This was the case in my study as well. While the interview data was collected, I made adaptations based on my notes of previous interviews, which led to a "repetitive interplay between the collection and analysis of data" (Bryman, 2008:539).

The ultimate goal of careful data coding needs to be a 'thick description', a narrative of what was found, that shows the full complexity and depth of what is going on (Holliday, 2010:99). Accordingly, I understood coding as a means to systematize the information derived from the interview data. This was necessary in order to dissect the different layers of meaning and importance contained in the students' responses. The main steps taken in this process were the following:

1. listening to all interviews + note taking + summaries of all interviews
2. straightforward content analysis + preliminary coding of all Session 1 interviews
3. recoding
4. content analysis + preliminary coding of Session 2 + 3 interviews
5. repeating procedure
6. establishing themes, compiling the codebook[42]
7. describing and interpreting the results according to different levels of codes (themes, topics)
8. writing and rewriting of Qualitative Findings Chapter
9. matching results with quantitative findings and discussing them in the light of relevant theory

This iterative approach follows Bryman's four stages of qualitative analysis (2008). It was suitable because it is a meticulous, systematic process and open for emerging themes. It meant that all interviews were examined, commented on and re-coded several times.

The actual process of thoroughly listening to the collected data started when most of the interviews were finished, i.e. when the participating classes had been tested three times and before the delayed vocabulary post-test was conducted. I decided to code manually with the help of word-processing software as the amount of data was relatively manageable and thus the coding process more personal.

The sound files were numbered and arranged in folders and split up in interviews conducted after Session 1, Session 2 and Session 3. Unusable data, for instance when an interviewee had not turned up for the successive session, were discarded. All interviews were listened to once or twice and preliminary notes (highlighted through circling, bolding, underlining) and short summaries of the content of each interview were written. Any interesting patterns were noted, any surprising features, whenever something was unexpected or puzzling, as well as any apparent inconsistencies were written down as comments. As these summaries formed the basis for deciding which of the interviews to choose for further analysis, and as they may already contain tentative ideas, concerns, and patterns, the "initial coding" process had already started then (Strauss and Corbin, 1998:102).

As there was a large number of interviews (36) it soon became obvious that it would not be possible or reasonable to transcribe all of them. Instead, I decided to focus on carefully examining those interviews that offered the most rich and relevant responses, i.e. those where responses were most rich in content and relevant for understanding the questions investigated

[42] See below for details.

in this project. The un-analysed interviews were too implausible to be considered. The literature suggests that it is fairly normal to disregard some interview data. For instance, Saldana (2013:15) pointed out that "only the most salient portions of the [data] corpus merit examination" and that the other half can be summarized or deleted, "leaving the primary half for intensive data analysis". However, transcribing as much as possible in order not to miss anything important was advisable.

Therefore, I decided to analyse the interviews conducted with five students, who were given the pseudonyms Runa, Jeppe, Silas, Fie, and Stina. These five interviewees were selected because they represented relevant 'cases', containing rich information. They addressed many interesting issues that were relevant for this study, but also contained what Pope et al. (2000:114) called "deviant or negative cases" and thus offered important new insights. The interviews conducted after the first testing session with the five interviewees were fully transcribed. For the transcription no software other than Word for Windows was used. The students' responses were subdivided into turns, i.e. a response to one question, rather than into numbered lines. This segmentation seemed appropriate, as the students' contributions were analysed at a macro-level rather than on a micro-level, i.e. important was what was being said at word level and most non-lingual expressions or gestures were not important and therefore not noted down, unless relevant. Pauses and interrupted words were included, however. In order to give a truthful account of what was being said, students' mistakes were not corrected. Table 8 shows how these transcriptions were then transferred into an interview grid consisting of columns for the numbered turns, for comments and for codes:

Table 9: Example Of An Interview Transcription Grid (Containing A Column For The Numbered Turns, Timing In Seconds, The Transcribed Turns, General Comments, And The Codes)

no	time	student/interviewer	comments	codes
1		I First of all, did you find the text difficult to read or …		
2	00:14	**S No, but that's usually if there if there is a word I don't know I just read further and then usually, ehm, it ehm makes, makes sense**	• *just reading on, ignoring unknown words and trusting that the (con)text can be understood anyway*	• strategy: ignoring • task approach

The first column numbers the spoken turn-units of both the interviewer and the interviewee. The timing of the recorded interview was noted down in seconds whenever needed in the second column. Column three contains the transcription of what was said during the interviews, with 'I' marking the contributions of the interviewer and 'S' indicating what the

students said. According to these grid entries, references to students' statements are read as: letter code for the interviewee, number of interview session, and turn of statement in the interview (e.g. Runa 1.33 = interviewee Runa, first interview, 33rd turn). In the fourth column, I noted down my own comments. Finally, in column five and six the students' statements were coded and recoded. The data were thus presented in a way that it is clear 1) what is being said and under which circumstances, 2) which thematic categories the interview statements could fall into, and 3) how what is being said could be interpreted. The later discussion of why the things that were being said are significant was based on this framework (Qualitative Findings, Discussion).

While the interviews were transcribed and entered into the interview grid, coding continued. No single method was applied, but different styles were adapted that seemed most useful for the data at hand (section 3.11.2). To use Saldana's (2013) terminology as an example, I applied "descriptive coding" (summarizing the basic topic of a passage of data, p. 70), "simultaneous coding" ("the application of two or more different codes to a single qualitative datum", p. 62), or also "magnitude coding" (adding alphanumeric codes to existing coded datum to indicate e.g. its intensity, frequency, direction, p. 58) at different stages of the coding process. More important than following specific coding techniques, was not to fragment the data, i.e. not to lose "sensitivity to the overarching narrative they are part of" (Bryman, 2008:553).

Another aspect that continually had to be considered was objectivity. Due to the interpretative nature of the analysis, particular caution was necessary. What one 'sees' in a transcription is inescapably selective and codes can thus become tainted with pre-supposed views. This can hardly be avoided, as "transcribers bring their own language ideology to the task. In other words, all transcription is representation, and there is no natural objective way in which talk can be written" (Roberts, 1997:168). Therefore, highly individualistic coding systems can be problematic, because they are difficult to compare across studies. External checks are suggested in the literature as a means to diminish this effect (Kim, 2008; Pope, 2000) and to establish coding reliability, i.e. system that can be used consistently across coders. Interrater reliability can be measured simply by using percentages[43] or Cohen's Kappa, which also accounts for chance agreements (Mackey & Gass, 2005:243). At the early stages of the coding process, I therefore consulted fellow PhD students as external appraisers of the coding system. By checking and comparing the categories established by each of the raters, they helped by using an approach similar to Cohen's Kappa.

[43] Measuring interrater reliability by percentage: "ratio of all coding agreements over the total number of coding decisions made by the coders" (Mackey & Gass, 2005:243).

As mentioned above, initial coding meant “breaking down qualitative data into discrete parts, closely examining them, and comparing them for similarities and differences” (Strauss & Corbin, 1998:102). This process is what Saldana (2013) calls ‘first cycle’ coding (initial coding and recoding), and which he distinguishes from ‘second cycle’ coding (classifying, prioritizing, integrating, synthesizing). I went through the text and coded everywhere interesting issues occurred[44]. When, for example, an interviewee described a specific vocabulary retention routine, I coded this as ‘vocabulary learning’ and ‘strategy’. Other statements were coded regarding specific ‘relationships’ (e.g. student/teacher/task) or because they revealed emotional aspects (e.g. test attitude, task criticism). Soon, major thematic ideas emerged in the data and I began linking chunks of data that represent the same phenomenon (“pattern coding”, Saldana, 2013:152). This process was then repeated. As a result of revisiting the data a number of times, some salient content categories emerged, linked to various data segments and it became obvious that some codes overlapped.

I compiled a ‘codebook’ as an attempt of finding a consistent fashion of coding (Saldana, 2013). For that a definition of the code was needed. I understood a code as a unit arrived at by trying to answer questions concerning a data item for instance ‘What does this item of data represent?’ or ‘Of what general category is this item of data an instance?’ (Bryman, 2008) or more generically as Saldana (2013) suggested: “What strikes me?” I made a list of codes and constantly added definitions of these codes. This helped with keeping focus on rigor and reliability. The codebook was a compilation of the interview statements according to the codes that emerged during the analysis. By using this as an instrument, areas of interest were identified and the occurrence and the specific content of the different codes were investigated. It made it easier to see how issues overlapped and changed across the interviews. During coding it turned out that few clear-cut categories emerged. Rather, what seemed relevant at one point sometimes later turned out to be insignificant and vice versa and sometimes interviewees’ statements were inconsistent or illogical. However, due to the explorative nature of the research questions, initially anything was relevant and often it were precisely these contradictions between and within interview data that later turned out to be significant. The codebook made it easier to detect these cases and to understand the opaque ones.

For the matter of time and feasibility, it was necessary to focus on codes that turned out to be particularly important regarding my research questions, for example noticing the TIEs, using the TIEs, attitudes towards the TIEs; and to disregard others that proved to be of fickle nature.

[44]I found valuable practical theoretical help at (http://onlineqda.hud.ac.uk/Intro_QDA/how_what_to_code.php).

Finally, from the 49 codes that were listed in the codebook at the point in time when the writing of the data analysis began[45], only those were chosen for further analysis that seemed most relevant with respect to the questions investigated in this study and in relation to the findings from the quantitative data. Still, even these codes were no absolute labels, but rather offered insight into different stages within the analytical process. They constantly changed, for instance when the codebook was compiled, items had to be relabelled and turns overlapped in terms of how they could be interpreted and therefore allocated to a category or code.

Several top-level codes ('themes') emerged, which during the compilation of the codebook and writing process turned out to contain numerous sub-codes, which were then examined in greater detail. These were, for instance, the theme "TIEs", which dealt with topics such as "how the TIEs were used", "noticing of TIEs" or "positive or critical attitude towards TIE use". Saldana (2013) suggests that the decision of whether codes would take more the form of lumping (big strokes coding) or splitting (smaller code units) should depend on the data units. I decided to strike a balance between the two and keep both the overarching issues ('themes') and related sub-topics ('codes') active throughout the analysis. This approach was also used in the final stages of writing up the qualitative results at 'theme' level.

During another round of coding, new themes and sub-codes were established with the help of the codebook. It became obvious that various statements and codes were puzzling and difficult to understand or classify. As mentioned above, a good way of handling critical cases and making elaborate decisions is to use interraters. However, while critical instances were discussed with other researchers and also in the supervision process, in the later stages of this study there was no systematic cooperation with interraters, mainly for time reasons. Therefore, all ambivalent and surprising cases were treated as genuine part of the data, which can lead to important insights, especially when triangulated with related findings from the quantitative data. At the same time, statements from interviews needed to be treated cautiously and could not be understood as an 'irrevocable truth'. I had to acknowledge, that responses might have been spontaneous rather than reflected, were contradictive across interview sessions, and that the interviewees were students interviewed in a school context and perhaps under pressure to perform well.

Writing up the data for the Qualitative Findings chapter and then for the Discussion was the final step of the analysis. Here, it was necessary to establish some sort of 'coding hierarchy', i.e. I had to decide which the most important codes were in relation to my research questions. In the Discussion Chapter, one guiding principle was to find codes which would be suitable to

[45] It should be noted that even after returning to the interview data several times, the number and labels of the chosen codes was preliminary and new codes appeared throughout the analytical process.

relate to findings from quantitative data. Even though quantizing or standardizing the qualitative data for statistical analysis was not necessary, it needed to be relatable and therefore comparable (Tashakkori & Teddlie 2010). There were, for example, many interview statements showing differences in focus (e.g. on form, on meaning). These were particularly relevant for interpretation in the Discussion.

Lastly, it is necessary to mention three matters. Firstly, the second part of the interviews, which regarded knowledge of specific target words, was not included in the data analysis. The data were not collected rigorously enough to be insightful and the students' statements could not be reasonably well linked to the quantitative data. Secondly, it was likely that the interviews had an impact on the long-term vocabulary retention results. Therefore the scores from the interviewed students were removed from the main data base. Thirdly, it seems that the alleged pressure to perform well in the VPT made some participants exaggerate and forget about their natural use of textual enhancements. Being interviewed may have amounted to even more pressure and may have led the interviewees to adapting their approach. This kind of behaviour is difficult to avoid, even if ethical considerations are applied and a relaxed atmosphere and a good rapport with the participants is established.

The following chapter presents the findings from the analysis of the quantitative data.

4. QUANTITATIVE FINDINGS

4.1. Introduction

This chapter presents outcomes of the statistical analysis of the quantitative data gathered in the main data collection and the delayed vocabulary post-test (VPT).

I investigated the dependent variable, vocabulary acquisition, with regards to how it was influenced by the independent variables Textual Input Enhancement (TIE), session, school, gender, text, language background, and in the case of the delayed post-test data, the variable time. The relevance of these variables is explained in the Methodology Chapter.

The dependent variable, vocabulary acquisition[46], was measured as points scored in the VPT ranging from 0 – 10. For reasons explained in the Methodology Chapter, it is treated as an ordinal variable. The independent variable TIE is regarded as a nominal variable[47], which measures the different forms of TIE at four levels representing the control group and the three enhancement types.

For reasons explained in the Methodology Chapter, only the results of vocabulary post-test parts B-D are reported, while those of test part A are disregarded. As a reminder, the set-up of the VPT is outlined again: There were four parts to the VPT and the different parts measured different types of vocabulary knowledge. The very first part (A), tested passive form recognition. The participants were asked to mark whether they had seen or knew a particular word in a table (Table Appendix 1). However, as was explained previously, this part of the test served merely as a primer for the following three parts and the results were not included in the analysis. In part B, the learners were tested on active form recognition. They were given a word definition and a choice of four different words and the option 'I don't know': The task was to choose the word that matched the definition (Figure Appendix 4). In part C of the test 'passive meaning recall' was tested. The students were presented with the target words and asked to find synonyms, equivalent words or to rephrase the meaning of the word (Figure Appendix 5). In the final part, D, passive form recognition was tested by the participants having to find the best matching synonym to the given target word from a list of four different words and the option 'I don't know' (Figure Appendix 6).

[46] The term 'vocabulary acquisition' is used as an umbrella term, which covers different facets of word knowledge. It does not imply 'full' knowledge, as it is not likely to be achieved after a singular exposure to a new word in a text.
[47] This is most appropriate because then all variables are regarded as being of the same type. See Methodology Chapter for details.

The analysis started with an examination of the VPT data gathered in Session 1 only, i.e. from an incidental vocabulary learning situation. The outcomes of the analysis are presented in section 4.3. The sections thereafter present the results based on analysis of data gathered in sessions 2 and 3 (sections 4.4 and 4.5). The entire quantitative data body was used as one sample when the data from all testing sessions was pooled (section 4.6). This allowed for a presentation and comparison of results from the three test sessions separately and then also for examining the results from all three sessions combined.

4.2. The Hypotheses

As outlined in the previous section, the two major variables are 1) the different forms of textual enhancement (TIEs) and their effect on 2) the performance in the VPTs (Vocabulary Acquisition). Accordingly, the following hypotheses were formulated.

The Null Hypothesis (H_0) claims that there is no difference between the different TIEs in their effect on vocabulary learning. Accordingly, there is no relationship between TIE and Vocabulary Acquisition. In view of that, three alternative research hypotheses were tested. According to the first, there are differences between the different TIEs regarding their effect on vocabulary acquisition from reading. However, it is a non-directional hypothesis as the nature of these differences is not clear, i.e. which form of textual enhancement may be beneficial or confusing, and therefore be detrimental for vocabulary acquisition, is not known. Likewise, the second alternative hypothesis is based on the notion that there are degrees to the effect of textual enhancement. It suggests that TIEs are beneficial for vocabulary acquisition, i.e. that more text manipulation leads to higher scores in tests measuring vocabulary acquisition. Finally, hypothesis 3 suggests that textual manipulation has a negative effect, as it may, for example, be distracting.

Which of these hypotheses holds, is examined in the following analysis.

4.3. Findings Session 1

The results reported in this section are based on the data collected in the first testing session. This was done to work first with data untainted from test awareness and to get an insight into whether familiarity with the test procedure in the last two test sessions (‘test familiarity’) had an impact on their VPT performance.

First, basic descriptive statistics are reported to offer insights into the nature of the data (Tables 10 and 11). Table 10 shows the distribution for the three test parts according to reading condition.

Table 10: Session 1 - Descriptive Statistics per Reading Condition, Parts B-D

	B				C				D			
	C 0	**C 1**	**C 2**	**C 3**	**C 0**	**C 1**	**C 2**	**C 3**	**C 0**	**C 1**	**C 2**	**C 3**
N	77	80	54	58	76	79	53	58	77	77	52	57
Mean	7.1	6	7.9	7.2	2.3	1.6	4.3	3.3	5.2	4.4	7.2	6.5
SD	2.6	2.9	2.0	2.5	2.3	1.9	2.4	2.5	2.5	2.6	1.9	2.5
Min	1	0	2	1	0	0	0	0	0	0	2	0
Max	10	10	10	10	9	9	10	9	10	10	10	10

In all three VPT parts, the highest mean scores were achieved when working with reading condition 2 (the glossed target words). Strikingly, the control group (reading an unenhanced text) outperformed the groups of learners reading texts in which the target words were bold printed only (condition 1). Table 11 below lists the scores for the three parts of the vocabulary post-test across the different reading conditions.

Table 11. Session 1 – Descriptive Statistics for Vocabulary Post-Test, Parts B-D

		VocScore B active form recognition	**VocScore C passive meaning recall**	**VocScore D passive meaning recognition**
	valid	269	266	263
	missing	0	3	6
Mean		6,94	2,69	5,62
Median		7,00	2,00	6,00
Mode		10	0	6
Std. Deviation		2,615	2,456	2,641
Variance		6,840	6,034	6,976
Skewness		-,726	,839	-,283
Std. Error of Skewness		,149	,149	,150
Range		10	10	10

Table 9 shows the results for Session 1 for VPT parts B, C and D. The slight decrease of the number of participants from B to D suggests that some participants did not complete the whole set of VPT exercises, possibly because they ran out of time.

VPT part B, which gauged active form recognition, led to the highest mean scores (M: 6,94 out of 10). In contrast, the scores for part C (passive meaning recall) are lowest. It might be surprising that the mode for part B is as high as 10. This could be seen as an indication of the fact that selecting the correct word form with a provided meaning is not as 'active' as Laufer et al.'s labelling suggests. Another reason for the high mode may be that many of the participants in this session worked on Text 1, which was – despite carefully selecting texts of

similar difficulty - according to some of the interviewees, easier to understand. This issue is addressed in section 4.6.3 and in the Discussion.

Differences in variance are relevant to regard, as equality of variance is assumed by some statistical instruments. The three variance values shown above are very similar. Whether or not there are significant differences can be accurately measured by the Levene's test. However, "[a] rule of thumb with equality of variance is that the largest SD of your groups should not be more than twice the smallest SD of your groups" (Lowie & Seton, 2012:93). As this is given for the Session 1 data, diversity in the variance does not present a problem.

As for skewness, it should be noted that the scores for part B and D are negatively skewed, with the values for B being more extreme than for part D. This indicates a rather easy test, because of the many occurrences at the high end of the distribution, i.e. many high scores and relatively few low scores (Salkind, 2012). However, for part C the results are positively skewed, so that it must be assumed that this part was rather difficult and only few participants recalled a high number of words. These observations suggest a non-normal distribution of the data, but are not surprising - it is often assumed that word recognition is less challenging than word recall (Alter & Oppenheimer, 2009; Laufer, et al. 2004; Laufer & Goldstein, 2004; Nation, 2001).

However, to choose the most suitable statistical instruments, the data distribution had to be investigated in more detail. Two tests of normality, the Kolmogorov-Smirnov and the Shapiro-Wilk test were conducted. While the Kolmogorov-Smirnov test is best used for sample sizes >50, the Shapiro-Wilk test is regarded as more reliable for smaller sample sizes (Lowie & Seton, 2012:92). Both assumed as the null hypothesis that the data were normally distributed. The results are reported in Table Appendix 3. As assessed by those two tests, the scores for VPT part B, C and D are not normally distributed ($p<.05$). As the VPT scores contain many values close to zero, are naturally limited (by 0 and 10) and as there are many tied values[48], this may not be surprising. However, as a consequence, in the analysis non-parametric tools are needed as parametric tests usually assume normal data distribution and non-parametric tools do not.

To choose the appropriate tools for the statistical analysis, the type of the variables had also to be taken into account. As a reminder, we are dealing with a discrete ordinal ratio scale dependent variable (the VPT scores) and nominal independent variables (e.g. TIE, session). In the first session, the participants were part of independent groups. The data collected in Session 1 were investigated with respect to whether there was a significant difference in

[48] *Tied values* occur when two or more observations are equal, whether the observations occur in the same sample or in different samples. In theory, non-parametric tests were developed for continuous distributions where the probability of a tie is zero. In practice, however, ties often occur. http://v8doc.sas.com/sashtml/stat/chap47/sect13.htm, accessed on 2/5, 2013; see also Field 2005:524.

vocabulary intake based on TIE-type. To test the null hypothesis, the suitable statistical instrument had to determine differences between two or more groups of unrelated (independent) cases on the dependent variable and to find out where the differences between the groups lie. The set–up was a between-subject design, where each participant worked with only one level of the independent variable, namely one text enhanced in one condition. Furthermore, the statistical instrument had to work with a nominal independent variable with four levels (control group, bold-printing, glossing, bold-printing + glossing) and an ordinal discrete ratio scale dependent variable (VPT scores). For non-normally distributed data the appropriate instrument is the Kruskal-Wallis H Test. This non-parametric test evaluates differences in medians among groups and "uses the order of your data from the lowest to the highest values to create ranks" (Lowie & Seton, 2012:63). A chi-square statistic is used to evaluate differences in mean ranks to assess the null hypothesis that the medians are equal across the groups.

SPSS software was used to calculate the test statistic, H (Equation Appendix 1). The results for all three parts of the VPT are shown in Table 12.

Table 12. Session 1 – Kruskal-Wallis H Test Results

	VocScore B active form recognition	VocScore C passive meaning recall	VocScore D passive form recognition
Total N	269	266	263
Test Statistic H	15.221	48.667	42.333
Degrees of freedom	3	3	3
Asymptotic Sig. (2-sided test)	.002	.000	.000

The results of the Kruskal-Wallis H test suggest that the H_0-hypothesis of equality of distribution must be rejected as significant group differences were found in all three parts of the VPT (part B: $H(3) = 15.221$, $p = .002$; part C: $H(3) = 48.667$, $p = .000$; part D: $H(3) = 42.333$, $p = .000$). However, before interpreting this finding, it is important to find out exactly which TIEs are related, and where the differences between them lie. The outcomes of the Kruskal-Wallis H test do not show this, which is why a post-hoc test was needed. SPSS uses the Mann-Whitney U test with the Bonferroni test correction as a post-hoc test to the Kruskal-Wallis H test. It is a widely used a posteriori instrument "to show precisely where relationships between subgroups lie when the numbers of cases in each group are not equal"[49] (Field, 2005:339f).

The details of the post-hoc results for the three VPT parts from Session 1 are shown in Tables 5, 6 and 7 in the Appendix. In Table 13, the results are summarized for better reading, only the significant group differences are shown.

[49] See also http://www.wellesley.edu/Psychology/Psych205/anova.html; accessed on 1/2, 2012.

Table 13. Session 1 - Significant Group Differences, Based On Kruskal-Wallis H Test And Mann-Whitney U Post-Hoc Test - Summarized

B – active form recognition					C – passive meaning recall					D – passive meaning recognition				
Test statistic [50]	Std. error	Std. test statistic	Sig.	Adj. sig.	Test statistic	Std. error	Std. test statistic	Sig.	Adj. sig.	Test statistic	Std. error	Std. test statistic	Sig.	Adj. sig.
bold-print (C1) – gloss (C2)					**(C0) unenhanced – gloss (C2)**					**(C0) unenhanced – gloss (C2)**				
-50.586	13.570	-3.728	.000	.001	-64.819	13.594	-4.768	.000	.000	-56.782	13.569	-4.185	.000	.000
					(C1) bold-print – gloss (C2)					**(C0) unenhanced – bold-print+gloss (C3)**				
					-87.794	13.487	-6.509	.000	.000	-35.688	13.209	-2.702	.007	.041
					(C1) bold-print – bold-print+gloss (C3)					**(C1) bold-print – gloss (C2)**				
					-56.078	13.135	-4.269	.000	.000	-79.827	13.569	-5.883	.000	.000
										(C1) bold-print – bold-print+gloss (C3)				
										-58.733	13.209	-4.447	.000	.000

Pairwise comparisons in the post-hoc test showed that the scores for VPT part B were significantly different between Condition 1 (bold-printing) and Condition 2 (glossing; U = -50.586, p = .001), but not compared to the control group. It is interesting to consider here what this table does not show, i.e. where no significant group differences could be found: Notably, for active form recognition (B), no significant differences emerged between the control group and enhancement types other than the pure glossing (C2). Similarly, across all three parts of the VPT the control group performed significantly different only in relation to glossing in the recall test C (U = -64.819, p = .000) and in VPT part D to glossing (U = -56.782, p = .000), and C3 (bold-printing + glossing) (U = -35.688, p = .041), but not at all in contrast to C1 (bold-printing). This suggests for initial vocabulary learning, that there is an effect only when target words are glossed or glossed and bold-printed, but not for bold printing only. It is surprising that for both, VPT part B and C, except for the above discussed cases, there are no significant differences between the control group and the other reading conditions. It could have been expected that a rather obtrusive form of text manipulation like the combination of bold-printing + glossing would have an effect on the initial retention of target words - at least in comparison to the control group. That this is not the case might lie in the nature of the enhancement types; it might have to do with text difficulty or also with test type. This issue is investigated further in the Discussion Chapter.

For part C, passive meaning recall, in particular and similar to the outcomes for part B, there are significant differences shown between bold-printing and glossing (U = -87.794, p = .000), and bold-printing and bold-printing+glossing combined (U = -56.078, p = .000). This means that the different types of textual enhancement have a differentiating effect in these cases. Different from the results for VPT part B, however, there is a significant difference

[50] Test statistic, Std. error = standard error, Std. test statistic = standard test statistic, Sig. = significance, Adj. sig. = adjusted significance

between the control group and the scores gathered from glossing only (Table Appendix 6). Glossing seems to have a beneficial effect on passive meaning recall. When comparing the unenhanced and the glossed enhancements, glossing lead to superior results.

For the last part of the VPT, where passive form-meaning recognition was tested, the results are somewhat different. Here the control group performed significantly different from both pure glossing and bold-printing and glossing combined. This supports the view that enhancing target words in a text does have an effect on how well these target words can be recognized initially. Such differences were also found between bold-printing and glossing (U = -79.827, p = .000), and between bold-printing and bold-printing+glossing (U = -58.733, p = .000; Table Appendix 7).

To be able to draw valid conclusions from these outcomes, it can be enlightening to investigate the size of the effects of the group differences. A p-value in itself does not give information about the magnitude of an effect, it merely reveals where differences are significant. Therefore, a complementary to the statistical significances measured here is needed to show the magnitude of the phenomenon. Effect size calculations provide such a measure. The calculation used here is shown and explained in Equation Appendix 2.

Detailed results of the effect size calculations for Session 1 can be found in Table Appendix 8. For the first testing session the calculations mainly found effects of 'medium' size, except for the difference VocScoreC $r_{C1\text{-}C2}$ (r = -.562, large effect) and VocScoreD $r_{C0\text{-}C3}$ (r = -.234, 'small' effect) and VocScoreD $r_{C1\text{-}C2}$ (r = -.500; large effect). Given that the data stem from an authentic classroom setting, having mainly 'medium' and even some 'large' effect sizes can be regarded as very satisfactory. These effect sizes allow for some confidence in drawing conclusions from the findings and will therefore be considered in the Discussion.

In summary, it was found that in the data collected in the first testing session, TIEs did really have a differentiating impact on VPT performance, which, however, was significant only in some cases. Even the scores achieved by the control group (C0) were not always significantly different from the ones achieved by the treatment groups. Large effect sizes were measured for the difference between C1 and C2 in VPT part C and between the same groups in VPT part D. In all other group differences, 'TIE' had a 'medium' effect on vocabulary uptake. Given these findings, for Session 1 the null hypothesis that the TIEs would not have an effect on vocabulary acquisition can be rejected.

In the following section, the analysis of the data collected in the second testing session is reported with respect to the same question, namely whether the TIEs had a significant effect on vocabulary acquisition as measured in the VPT.

4.4. Results Session 2

Table 14 shows the descriptive statistics for vocabulary post-test parts B-D according to each reading condition used in session 2.

Table 14: Session 2 - Descriptive Statistics per Reading Condition, Parts B-D

	B				C				D			
	C 0	**C 1**	**C 2**	**C 3**	**C 0**	**C 1**	**C 2**	**C 3**	**C 0**	**C 1**	**C 2**	**C 3**
N	42	77	45	67	41	77	45	67	41	77	45	66
Mean	6.4	7	6.2	8	1.7	2.4	1.8	4.2	5.3	5.4	5	6.6
SD	2.4	2.5	3	2.5	1.8	2.2	2	2.8	2.0	2.6	2	2.5
Min	1	2	0	0	0	0	0	0	0	0	1	0
Max	10	10	10	10	9	8	9	10	10	10	10	10

Different from the results reported for the first reading and testing session (Table 10), this time the highest mean scores were found for reading condition 3 (bold-printed and glossed target words) in all three parts of the vocabulary post-test. Different from the session 1 results is also, that the control group scored lower than condition 1, but for part B and D, higher than condition 2 (gloss only).

Table 15 shows details of the descriptive statistics based on the data collected in the vocabulary post-test in the second testing session.

Table 15. Session 2 – Descriptive Statistics for Vocabulary Post-Test Scores, Parts B-D

		VocScore B active form recognition	**VocScore C Passive meaning recall**	**VocScore D passive form recognition**
	Valid	231	230	229[51]
	Missing	0	1	2
Mean		7,03	2,69	5,63
Median		8,00	2,00	6,00
Mode		9	1	6
Std. Deviation		2,660	2,510	2,458
Variance		7,073	6,302	6,040
Skewness		-,700	,951	-,012
Std. Error of Skewness		,160	,160	,161
Range		10	10	10

The values are almost identical to those of Session 1. The variances are slightly more varied, but still sufficiently similar. Again, the scores for both VPTs part B and D are negatively skewed, and those for part C positively skewed. It must therefore be assumed that the data

[51] Here again, the number of participants decreases slightly, because some of the students were either too slow or to demotivated to finish off all parts of the vocabulary post-test.

are not normally distributed. The two tests of normality confirmed this (Table Appendix 9). As in Session 1, the significance levels show that the data were not normally distributed. Therefore it was necessary to continue working with non-parametric models to analyse the data.

In this session, the same students were tested again. But this was not considered at this point, so that sample was treated as consisting of 'independent' groups. Thus, the potential influence of test familiarity or test fatigue was disregarded. These issues were, however, taken into account in the analysis of the pooled data.

Again the null hypothesis assumed that the different TIEs would not make a difference for vocabulary learning from reading. The Kruskal-Wallis H test was used to explore possible differences between groups and subsequently a Mann-Whitney U test with Bonferroni correction was run to find out where these differences lay. Table 16 lists the results split up for the three parts of the VPT.

Table 16. Session 2 - Kruskal-Wallis H Test Results

	VocScore B active recognition	VocScore C passive recall	VocScore D passive recognition
Total N	231	230	229
Test Statistic	20.858	33.831	16.442
Degrees of freedom	3	3	3
Asymptotic Sig. (2-sided test)	.000	.000	.001

Table 16 shows that the hypothesis of equality of distribution must be rejected for all three parts of the VPT: part B, $H(3) = 20.858$, $p = .000$, part C, $H(3) = 33.831$, $p = .000$, and part D, $H(3) = 16.442$, $p = .001$. A summary of the post-hoc test results, presenting only the significant group differences, is shown in Table 17 (see Tables Appendix 11 – 13).

Table 17. Session 2 - Significant Group Differences Based On Post-Hoc To Kruskal Wallis H Test And Mann- Whitney U Post-Hoc Test - Summarized

B – active form recognition					C – passive meaning recall					D – passive meaning recognition				
Test statistic	Std. error	Std. test statistic	Sig.	Adj. sig.	Test statistic	Std. error	Std. test statistic	Sig.	Adj. sig.	Test statistic	Std. error	Std. test statistic	Sig.	Adj. sig.
unenhanced (C0) – bold-print + gloss (C3)					**unenhanced (C0) – bold-print + gloss (C3)**					**unenhanced (C0) – bold-print + gloss (C3)**				
-50.08	12.99	-3.85	.00	.00	-63.10	13.02	-4.84	.00	.00	-34.67	13.06	-2.65	.00	.048
bold-print (C1) – bold-print + gloss (C3)					**bold-print (C1) – bold-print + gloss (C3)**					**bold-print (C1)– bold-print + gloss (C3)**				
-32.27	11.03	-2.92	.00	.02	-41.19	10.97	-3.75	.00	.00	-33.78	11.02	-3.06	.00	.048
gloss (C2) – bold-print + gloss (C3)					**gloss(C2) – bold-print + gloss (C3)**					**gloss (C2) – bold-print + gloss (C3)**				
-47.32	12.72	-3.71	.00	.00	-60.52	12.65	-4.78	.00	.00	-46.38	12.70	-3.65	.00	.002

When considering these significantly different distributions, it is necessary to remember that in comparison to Session 1, this time the participants were going through the reading/testing

process for the second time and might have been aware of the upcoming VPT. Therefore, a purely incidental learning situation cannot be assumed. This is considered in the Discussion.

Considering the significant group differences found between the groups tested in Session 2, there are significant differences between the same groups across all three parts of the VPT, namely in all three cases between the control group and bold-printing+glossing (B: U = -50.082, p = .001; C: U = -63.100, p = .000; D: U = -34.670, p = .048), between bold-printing and bold-printing+glossing (B: U = -32.270, p = .021; C: U = -41.192, p = .001; D: U = -33.786, p = .048), and between glossing and bold-printing+glossing (B: U = -47.327, p = .001; C: U = -60.535, p = .000; D: U = -46.386, p = .002). Furthermore, it is striking that the results for the control group are only significantly different from the combination of bold-printing + glossing, but not from the other two types of TIEs. This suggests that compared to dealing with an unenhanced text, only a strong form of TIE (e.g. bold-printing + glossing) seems to have a significant impact on vocabulary retention. The 'weaker'/less intrusive forms did not have such a strong influence. This is further underlined by the fact that whenever the group differences were found to be significant, it was in comparison to where the target words were enhanced twofold.

In summary, for Session 2, significant group differences were found between the following groups: control group (C0) - bold-printing+glossing (C3), bold-printing (C1) - bold-printing+glossing (C3), and glossing (C2) - bold-printing+glossing (C3) across all three parts of the VPT. For the significant group differences in Session 2, mainly effects of medium size were found, except for three cases of small effects (part B: $r_{C1\text{-}C3}$ = -.243, part D: $r_{C0\text{-}C3}$ = -.254; $r_{C1\text{-}C3}$ = -.255; Table Appendix 14). This suggests that TIE had a considerable impact on the participants' performance in the VPT. Therefore, the null hypothesis, assuming no significant effect of TIEs on vocabulary acquisition, must be rejected again.

Below, I present the results based on the statistical analysis of the data collected in the final testing session.

4.5. Results Session 3

Below (Table 18) the mean scores for all three vocabulary post-test parts are reported according to the three different reading conditions and the outcomes of the control group.

Table 18. Session 3 - Descriptive Statistics per Reading Condition, Parts B-D

	B				C				D			
	C 0	**C 1**	**C 2**	**C 3**	**C 0**	**C 1**	**C 2**	**C 3**	**C 0**	**C 1**	**C 2**	**C 3**
N	47	19	64	58	47	19	62	58	47	19	64	58
Mean	6.9	8	7.9	7.9	1.7	3.4	3	2.9	4.4	5.8	5.3	5
SD	2.7	2.4	2.7	2.3	1.7	2.8	2.5	2.1	2.5	2.1	2.7	2.4
Min	1	3	0	2	0	0	0	0	0	2	0	1
Max	10	10	10	10	7	9	9	10	10	10	10	10

The results are different from the first (Table 10) and second (Table 14) reading and testing sessions. It is striking here, that the highest scores were found for reading condition 1 (bold-printed target words) across the three VPT parts. However, the relatively low number of participants in this part of the sample needs to be noted. The control group achieved the lowest scores.

The descriptive statistics of the data collected in Session 3 are presented in Table 19:

Table 19. Session 3 - Descriptive Statistics for Vocabulary Post-Test, Parts B-D

		VocScore B active form recognition	**VocScore C passive meaning recall**	**VocScore D passive meaning recognition**
N	**valid**	188	186	188[52]
	missing	0	2	0
Mean		7,57	2,70	5,06
Median		8,00	2,00	5,00
Mode		10	2	5
Std. Deviation		2,546	2,308	2,476
Variance		6,481	5,325	6,130
Skewness		-,851	,891	-,046
Std. Error of Skewness		,177	,178	,177
Range		10	10	10

Just as for Session 1 and 2, the numbers in Table 15 are only interesting in that the skewness renders a normal data distribution shape unlikely. Therefore, the data distribution had also to be checked to find the most appropriate statistical tools for data analysis. Accordingly, the Shapiro-Wilk and the Kolmogorov-Smirnov test were applied. Table Appendix 15 shows the detailed test results. They confirm that the data are significantly non-normally distributed. Therefore only statistical models not assuming normal distribution were applied in the data analysis.

[52] The number of participants had decreased, as two classes turned out to be only available for two testing sessions.

For the data collected in this final testing session, the null hypothesis assumes that there is no differential effect on vocabulary intake based on the impact of the TIEs.

To detect group differences the Kruskal-Wallis H test was applied. The same groups of students were tested who had also participated in testing Session 1 and 2. Still, the sample was treated as consisting of 'independent' groups as only the Session 3 data was of interest. Potentially influential factors that might be connected to the fact that the students were tested repeatedly (test familiarity and test fatigue) were disregarded again. They were, however, considered later on in the analysis (see section 4.6.3). Table 20 shows the results of the Kruskal-Wallis H test.

Table 20. Session 3 - Kruskal-Wallis H Test Results

	VocScore B active form recognition	VocScore C passive meaning recall	VocScore D passive meaning recognition
Total N	188	186	188
Test Statistic	6.080	12.39	5.666
Degrees of freedom	3	3	3
Asymptotic Sig. (2-sided test)	.108	.006	.129

The results indicate that significant group differences were only found between the scores of VPT part C ($H(3) = 12.39$, $p = .006$), the meaning recall test, but not for part B ($H(3) = 6.080$, $p = .108$) and D ($H(3) = 5.666$, $p = .129$). Therefore, the post-hoc test. A Mann-Whitney U with Bonferroni correction was only run for the scores of part C (Table Appendix 16). For better reading, the significant group differences are summarized in Table 21.

Table 21. Session 3 - Significant Group Differences Based On Post-Hoc to Kruskal Wallis H Test and Mann- Whitney U Post-Hoc Test – Summarized

C – passive meaning recall				
Test statistic	**Std. error**	**Std. test statistic**	**Sig.**	**Adj. sig.**
Control group (C0) – bold-print (C1)				
-39.879	14.449	-2.760	.006	.035
Control group (C0) – gloss (C2)				
-28.946	10.279	-2.816	.005	.029
Control group (C0) – bold-print+gloss (C3)				
-29.787	10.431	-2.856	.004	.026

Table 17 indicates significant differences between the control group and the bold-printing group ($U = -39.879$, $p = .035$), the control group and the glossing group ($U = -28.946$, $p = .029$), and between the control group and the bold-printing+glossing group ($U = -29.787$, $p = .026$). These findings suggest that in the last round of testing the TIEs had no significant impact

on active or passive form recognition (part B and D), but only on meaning recall (C). It may be that learners focussed on the target words to an extent that the effects of different enhancement forms were overridden. While this intentional reading approach worked well for the parts of the test that tested recognition, it was not sufficient for the more difficult recall part of the test. However, this would not explain why the participants working with enhanced forms did not score higher than those exposed to the control group in the VPT parts that assessed recognition (B and D). The details need to be investigated further and are explored in the Discussion. The results of part C are noteworthy because it was the first time in this session that the control group produced outcomes that were significantly different from those of all three treatments groups. This suggests that a recall task is difficult even when learners are familiar with the test set-up and that any form of enhancement can be used in order to gain better results than when no enhancements are provided. The fact that the learners were probably reading intentionally is likely to have helped, even those participants who worked with bold-printed target words.

The calculated effect size measures fell into the category of 'medium effects' for the group difference found between the control group and the bold-printing group (VocScoreC $r_{C0\text{-}C1}$ = -.346), and into the category 'small' effect for the two other found significant group differences (VocScoreC $r_{C0\text{-}C2}$ = -.266, VocScoreC $r_{C0\text{-}C3}$ = -.280, Table Appendix 17).

To conclude, it is important to point out that this time, significant group differences were only found in the scores for active vocabulary recall (VPT part C). Most strikingly, the control group performed significantly different from all three of the treatment groups. Considering these outcomes, for the data collected in Session 3, the null hypothesis that the impact of the different TIEs would not lead to a measurable significant effect on vocabulary analysis holds for VPT parts B and D (word recognition), but must be rejected for the passive recall task, i.e. part C (production).

4.6. Pooled Data

The following section addresses the question of whether the group differences found so far also appear when the scores from the three testing sessions are pooled, i.e. when the data that were split up according to session are brought together into one data 'pool'. Considering the scores from all these three sessions and all participants together, i.e. pooling the data, takes account of the fact that the results are interconnected as the same participants were tested repeatedly. It has the advantage that a large number of data points are available for investigating the effect of each the TIE-types. Thus, conclusions are drawn from scores from

a larger number of participants, which makes it easier to generalize findings. Section 4.6.2 shows findings regarding the size of the correlation between the variables TIE and vocabulary acquisition based on data collected in all three testing sessions. Section 4.6.3 reports the findings from investigations of the impact of further measured independent variables. It also presents the findings regarding the effect of the TIEs on different types of word knowledge.

4.6.1. Findings From Pooled Data

Across the three testing sessions, the participants were tested on all three TIEs. This gave the opportunity to investigate whether or not there was a difference in vocabulary test scores based on TIEs when the data from the three sessions are pooled. The data from the three sessions was pooled by treating all VPT scores together as one sample. They were treated as data falling into the four TIE types (Condition 1-3 and the control group) as distinguishing categories, while the potential impact of other factors was disregarded. Thus, it was possible to investigate whether outcomes of previous calculations could be confirmed or new insights could be gained. Consequently, the null hypothesis tested here is formulated in the same way as the previous ones; it anticipates that the TIEs have no differential impact on vocabulary uptake from the reading tasks, even when the data from the three sessions are pooled.

Unlike in the previous sections, the set-up is now a within-subject design, because the subjects were measured repeatedly on the same dependent variable but worked with different TIE-types. Each participant provided data for all three testing sessions, so that by the end of Session 3 all participants had worked with all three TIE types. The Friedman test is suitable to investigate group differences between an ordinal dependent variable and a three-level nominal independent variable (Larson-Hall, 2010)**.** The test results, shown in Table 22, reveal significant group differences between all three parts of the VPT.

Table 22. Session 1-3 Pooled Data, Significant Levels Friedman Test

	VocScore B active form recognition	VocScore C passive meaning recall	VocScore D passive meaning recognition
Total N	146	151	156
Test Statistic	14.111	32.684	28.235
Degrees of freedom	3	3	3
Asymptotic Sig. (2-sided test)	.003	.000	.000

The Wilcoxon W test was applied as a post-hoc test to find out where the significant differences lie. The findings are presented in Table 23 (for detailed results see Tables Appendix 19 – 21).

Table 23. Session 1-3 Pooled Data, Significant Group Differences Based On Post-Hoc W Test To Friedman's Test - Summarized

B – active form recognition					C – passive meaning recall					D – passive meaning recognition				
Test statistic	Std. error	Std. test statistic	Sig.	Adj. sig.	Test statistic	Std. error	Std. test statistic	Sig.	Adj. sig.	Test statistic	Std. error	Std. test statistic	Sig.	Adj. sig.
bold-print (C1) – bold-print+gloss (C3)					unenhanced (C0) – gloss (C2)					unenhanced (C0) – bold-print+gloss (C3)				
-.514	.151	-3.400	.001	.004	-.411	.149	-2.764	.006	.034	-.612	.146	-4.188	.000	.000
					unenhanced (C0) – bold-print+gloss (C3)					bold-print (C1) – bold-print+gloss (C3)				
					-.626	.149	-4.212	.000	.000	-.660	.146	-4.517	.000	.000
					bold-print (C1) – gloss (C2)									
					-.480	.149	-3.232	.006	.007					
					bold-print (C1) – bold-print+gloss (C3)									
					-.695	.149	-4.680	.000	.000					

When comparing these outcomes to the findings which investigated the scores from each session separately, no clear pattern is visible. None of the results are the same across all three separately investigated sessions or when considering the pooled data. However, a clear tendency emerged in the fact that working with bold-printed+glossed texts made it particularly likely to yield results significantly different from working with a text enhanced with any of the other TIEs. In other words, reading a text in which target words are enhanced twofold is more likely to have a significant effect on vocabulary acquisition than reading a text in which target words are not at all enhanced or in a less invasive form. It is also observable that for all three parts of the VPT a significant difference was measured between the scores of bold-printing and bold-printing+glossing (B: W = -.514, p = .004; C: W = -.695, p = .000; W = -.660, p = .000), i.e. between the 'weakest'/'least invasive' and the 'strongest'/'most invasive' form of enhancement. In the passive recognition task, a significant group difference was also found between the control group and the bold-printing+glossing group (W = -.612, p = .000). Again, it was for the recall part of the vocabulary test (C) that the largest number of group differences was calculated ($W_{C0\text{-}C2}$ = -.411, p = .034; $W_{C0\text{-}C3}$ = -.626; p = .000; $W_{C1\text{-}C2}$ = -.480, p = .007). In the same way as when regarding the results from the testing sessions separately, this suggests that reading enhanced texts had the clearest impact on form – meaning recall.

This time, the measured effect sizes fell into the category 'small effect' (see Table Appendix 22).

According to the results presented in this section, the null hypothesis, that there would not be a measurable differentiating effect of the TIEs on vocabulary learning when pooling the data, can be rejected. Pooling the data showed results that confirmed an overall tendency found in the outcomes from the three sessions, namely that working with bold-printed+glossed texts (C3) tends to yield significantly better results especially when compared to non-glossed TIEs or the unenhanced text of the control group. This suggests that more invasive forms of enhancement support vocabulary acquisition successfully. Several features can explain this.

One is that this combination of two attention-drawing techniques is likely to guide a reader's focus on both form and meaning, so that it is easier for a learner to link the new information about the word to already existing related knowledge (e.g. in a related lexical field). Moreover, the target word's salience and accordingly a learner's noticing of the word is greater than for simple enhancements such as bold-printing or glossing only (see Izumi, 2002 for similar finding and interpretation). Furthermore, combining a typographical and a lexical form of enhancement makes the purpose of the enhancements obvious for the reader, as there is a clear connection between a highlighted target word and the given explanations in the margin. Watanabe (19997:300) found this "clarity of connection between the explanations and [the] words to be explained" vital for learning words initially and for retaining them later. In addition, the typographic connection is likely to urge the reader to decontextualize, check, and process the target words. O´Donnell (2012:558) assumed that such clear, albeit obtrusive enhancements have the additional benefit of forcing the learner to reread the target words and their context, "presumably in order to maintain the storyline".

However, examining significant group differences in the pooled data showed no clear results regarding the TIEs' impact on the two assessed recognition types. The same was true for the results yielded from the data split up according to session. This is in line with Laufer et al. (2004), whose comparison between passive and active recognition showed no clear distinction either. In my analysis it is striking, however, that it was again condition three that affected test outcomes significantly.

So far, the case that different TIEs can have an effect on the outcome of the VPTs has been established. However, this in itself does not sufficiently allow for reliable interpretation. To achieve this, it is relevant also to examine the magnitude of the effects. This aspect is investigated in the following section.

4.6.2. Correlation Size Between Enhancement Type And Vocabulary Acquisition

So far, it has been established that the two variables TIE and vocabulary acquisition are related. To better understand the nature of this relationship and the underlying learning processes, it is relevant to gauge how strongly they are related. Therefore, this section shows to which extent TIE accounts for the outcome of the VPT score.

The correlation coefficient, i.e. a measure of the strength of the relationship between the two variables vocabulary intake and TIE, was calculated (Field, 2005) for the pooled data from the three testing sessions. 'Kendall's tau' (τ) was chosen as the suitable non-parametric correlation test, as it is regarded as an accurate correlation gauge (Field, 2005; Lowie & Seton,

2012; Newson, 2002). It is especially suitable for data with a large number of tied ranks and also with particularly small data sets and superior to Spearman's correlation *r* as a rank-based measure of correlation (Newson, 2002). The Kendall's Tau rank correlation coefficient assesses statistical associations based on the ranks of the data. Ranking data is carried out on the variables that are separately put in order and numbered. The magnitude of the Kendall's tau (τ) correlation coefficient determines the strength of the correlation. Correlation coefficients take values between minus one and plus one. The positive correlation signifies that the ranks of both the variables are increasing, and vice versa. The results of the test are reported in Table 24.

Table 24. Kendall's Tau Correlation Coefficient between TIE and VPT Parts B-D, Based On Pooled Data

			TIE	B active form recognition	C passive meaning recall	D passive meaning recognition
Kendall's tau		correlation coefficient		.570	.214**	.203**
	TIE	Sig. (2-tailed)		.241	.000	.000
		N		269	266	263
		correlation coefficient	.057		.507**	.515**
	B active form recognition	Sig. (2-tailed)	.241		.000	.000
		N	269		266	263
		correlation coefficient	.214**	.507**		.611**
	C passive meaning recall	Sig. (2-tailed)	.000	.000		.000
		N	266	266		261
		correlation coefficient	.203**	.515**	.611**	
	D passive meaning recognition	Sig. (2-tailed)	.000	.000	.000	
		N	263	263	261	

**. Correlation is significant at the 0.01 level (2-tailed).

Table 20 shows that a positive significant correlation with the variable TIE was found between the outcomes of part C and D, which suggests that TIE and vocabulary acquisition are related in these two cases, but not for part B. The Kendall's tau correlation coefficient suggests a medium to small correlation in both cases (part C: τ = .214, *p* = .000; part D: τ = .203, *p* = .000), which is a typical and satisfactory result for this non-parametric test[53]. The positive relationship suggests that the more obtrusive form of enhancement leads to higher number of recalled (part C) or recognized (part D) words. This finding proposes again that combining bold-printing and glossing (C3) lead to higher scores in the VPT than the other types.

Correlations were also found between different parts of the VPT. These findings indicate that an increase in one variable may also make an increase in the other variable likely. This would mean that those students who performed well, i.e. with a high score, in VPT part B were significantly more likely to also achieve high results in part C (τ = .507, *p* < .01) and also D (τ = .515, *p*<.01). Those students, who did well in the recognition task, also achieved high scores

[53] https://statistics.laerd.com/premium/sroc/spearmans-rank-order-correlation-in-spss-8.php, accessed on 14/9, 2015.

in the recall task, which suggests that the different vocabulary knowledge types are linked to proficiency. This relationship will be considered in the Discussion.

However, caution is needed, firstly because generally no clear-cut conclusions about causality can be drawn from a correlation coefficient (Field, 2005)**.** This is especially true as the data were gathered in a classroom, i.e. a multifactorial environment. There are several more reasons for being careful with interpreting the outcomes. One is that it is, albeit not impossible[54], somewhat controversial to look at these types of data (ordinal and categorical) in terms of a correlational relationship, because a linear or monotonic relationship between these two variables is questionable (Heiman, 2014; Rumsey, 2011). Another is that allocating a higher or lower value to 'TIE' presupposes that my claim of more or less invasive forms of TIE is accepted. Otherwise, the values are – statistically speaking – arbitrary, and typical correlations are not possible[55]. Similarly, as the VPT scores are discrete ordinal data points that do not vary on a continuous scale, one cannot fully conceive of an 'increase' of the value caused by the impact of working with a specific TIE. Statistically this is the case, but with respect to vocabulary knowledge research showed that knowledge is a fluid entity that can be imagined on a continuum (see Literature Review). This is only to some extent assessed in my test.

Therefore, the results of this correlation calculation represent an incident where the quantitative analysis is of only limited value for a better understanding of the data. Still, as already discussed in the introduction to this chapter, the nature of the dependent variable vocabulary acquisition presents a borderline case as it is ordinal, discrete, and measured on a ratio scale, but with eleven scale points also has continuous features. This applies also to the independent variable TIE: there is an implicit order in TIE, as it can be claimed that bold-printing (or the control group, for that matter) enhances the target words to a lesser degree than glossing and to an even lesser degree than a combination of the two. In addition to that, previous research, which had assumed degrees of enhancement found evidence that such a gradation of effects on vocabulary acquisition does exist (Watanabe, 1997). Therefore, conducting the correlation analysis and considering its outcomes was important for getting some insight into the magnitude and nature of the relationship between the VPT outcomes and the TIEs. The correlation results are taken up again in the Discussion.

To sum up, the Kendall's tau calculations revealed a significantly large correlation between TIE and VPT-performance for part C and D of the test. However, these results have to be

[54] See also http://stats.stackexchange.com/questions/23938/how-to-correlate-ordinal-and-nominal-variables-in-spss about the phi-coefficient, accessed on 7/9, 2015.
[55]For a detailed discussion of this matter see http://spssx-discussion.1045642.n5.nabble.com/correlation-and-regression-for-ordinal-and-nominal-dependent-td5471812.html, accessed at 4/6, 2013.

treated with caution. Providing more information in form of other coefficients, such as the regression coefficients below, will add to the robustness of the conclusions.

4.6.3. Impact Of Other Measured Factors And Results According To Vocabulary Knowledge Type

The third research question inquires which factors other than TIE measurably contribute to the acquisition of vocabulary from reading enhanced texts. This question arises because many other relevant factors interact simultaneously when TIE-learning is investigated in a real-life classroom situation. Many of those were difficult to measure in my research set-up, but others could be analysed to achieve a comprehensive understanding of the learning situation. The findings presented here are based on calculations that integrate these factors into a statistical model which sheds light on the impact these might have on the outcome of the VPTs. The results also provide information for the second research question, whether input enhancement does have a differential effect on different types of word knowledge.

The data collection included several potential predictor variables, related to the circumstances of the test outcomes, such as testing session, the participants' school[56], the text, their gender, and their language background. The null hypothesis was that none of these independent variables would have an effect on vocabulary acquisition, meaning that all of the regression coefficients were equal to zero.

To investigate this hypothesis, a regression model was needed that would measure the relation and interaction between the multiple independent nominal variables (TIE, session, gender, class, school, language background and text) that may have had an effect on the dependent variable, (vocabulary acquisition), and could reveal which of these factors may predict the outcome of the dependent variable.

A regression model makes it possible to capture the simultaneous correlation between related factors and can show how different or related variables are. It can also, to a certain extent, indicate which variable contributes most to an outcome (Lowie & Seton, 2012). If there is a relationship between the factors and the dependent variable, then a regression allows conclusions about the existence of such a relationship and also makes it possible to predict the value of one variable based on the other variable (Lowie & Seton, 2012; Rasinger, 2008)**.**

One presupposition for a simple linear regression is a linear relationship (Equation Appendix 3). This is not given in the case of my data. Therefore, a model was needed to transform the data into a log, so that linearity was not one of the required underlying assumptions. Furthermore, the suitable regression model needed to allow the prediction of

[56] Also the class the students attend might play a role. It was, however too difficult to include this factor in a regression model, as – with overall 12 participating classes – too many dummy variables would have had to be included into the model, rendering the result too unreliable.

the test outcome based not only on one independent variable (e.g. on TIE), but also on the other factors. As there was more than one explanatory variable, a multiple regression model was needed, where each predictor has its own coefficient.

Therefore, an ordinal multiple logistic regression was used to work with the ordinal dependent variable (Larson-Hall, 2010; Orme & Combs-Orme, 2009; see Equation Appendix 4). The model estimates from a logistic regression are maximum likelihood estimates arrived at through an iterative process. In this process, ordinal logistic regression works with odds - it calculates the probability of an event occurring against the probability of it not occurring[57].

There are two further fundamental assumptions to ordinal logistic regression. One is the avoidance of multi-collinearity and the other is the assumption of proportional odds. In my study the first, the assumption of no multi-collinearity, i.e. that one (or more) of the independent variables is highly correlated with the other independent variables in the regression equation, was assured by a test that is automatically included in the ordinal logistic regression calculation in Stata (Cohen, Cohen, & Stephen, 2003). Also, proportional odds had to be verified (Long & Freese, 2006; Orme & Combs-Orme, 2009). Whether this assumption was met was assessed by a full L2 Log Likelihood ratio test using Stata software. The results are shown in Table 25.

Table 25. **Parallel Lines Test - Approximate Likelihood-Ratio Test of Proportionality of Odds across Response Categories, Results for Vocabulary Post-Test Parts B, C and D**

L2 Log Likelihood for B active form recognition	L2 Log Likelihood for C passive meaning recall	L2 Log Likelihood for D passive meaning recognition
chi2(99)[58] = 116.38	chi2(94) = 149.28	chi2(98) = 145.56
Prob > chi2 = 0.1120	Prob > chi2 = 0.0002	Prob > chi2 = 0.0013

According to these results, the assumption of proportional odds was met only by VPT part B, with $x^2 = 116.38$, $p = .1120$. In the case of VPT part C and D, however, the results show a violation of the parallel line assumption. This is often the case, as the test is highly conservative and reacts to even the smallest differences in the odds[59] (O'Connell, 2006). Therefore, its results are fairly commonly ignored (Long & Freese, 2006; Norusis, 2012). However, in order not to take any chances, other models (multinomial logistic regression, default generalized logistic regression, partial proportional odds model, see Williams, 2006) were tried as

[57]Ordinal logistic regression in SPSS. https://statistics.laerd.com/premium/olr/ordinal-logistic-regression-in-spss-21.php, accessed on 1/9, 2013.
[58] Chi2 (x) = chi-squared (degrees of freedom), tests that at least one of the predictors' regression coefficient is not equal to zero in the model; Prob > chi2 = this is the probability of obtaining this chi-square statistic (31.56) if there is in fact no effect of the predictor variables
[59] See also ESRC National Centre for Research Methods. http://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/modules/mod5/9/

potentially suitable alternatives. As they were found to be either unsuitable for covering all the factors under scrutiny here, or to not be applicable with the nature of the different variables (ordinal and categorical), the ordinal logistic regression was still considered to be the best model to investigate the data. The data had to be prepared by defining dummy variables, where after the regression was calculated by using Stata software and according to the stochastic model including the variables under investigation (Equation Appendix 5).

In the following section, the results of this ordinal regression are reported.

A regression aims to discover whether the coefficients on the independent variables are really different from zero - that the independent variables are having a genuine effect on the dependent variable - or if alternatively any apparent differences from zero are just due to random chance[60]. Therefore, only the parameter estimates showing a significant outcome for part B, C and D of the VPT are summarized in Table 22, because it is usually not appropriate to report a coefficient, if the test statistic for the regression coefficient is not statistically significant. However, the details of all the results of the ordinal logistic regression can be found in the appendix (Tables Appendix 23, 25, and 27).

[60] http://dss.princeton.edu/online_help/analysis/interpreting_regression.htm, accessed on 2/8, 2013.

Table 26. Summary of Significant Predictor Variables, Based On the Ordinal Logistic Regression

B – active form recognition						C – passive meaning recall						D – passive meaning recognition					
Coef	**Std err**	**z**	**P>z**	**Min conf**	**Max conf**	**Coef**	**Std err**	**z**	**P>z**	**Min conf**	**Max conf**	**Coef**[61]	**Std err**	**z**	**P>z**	**Min conf**	**Max conf**
glossing (C2)						**glossing (C2)**						**glossing (C2)**					
.39	.19	2.01	.04	.01	.79	.83	.20	4.12	.00	.43	1.22	.55	.19	2.82	.01	.17	.93
bold-printing+glossing (C3)						**bold-printing+glossing (C3)**						**bold-printing+glossing (C3)**					
.54	.19	2.78	.01	.16	.92	1.09	.19	5.57	.00	.71	1.49	.72	.19	3.70	.00	.34	1.09
Session 3						**Session 3**						**School 2**					
.88	.21	4.23	.00	.47	1.28	.49	.19	2.45	.01	.09	.88	-.39	.17	-2.21	.03	-.72	-.04
School 2						**School 2**						**Language 2**					
-.46	.18	-2.63	.01	-.81	-.12	-.63	.18	-3.60	.00	-.99	-.29	-.68	.21	-3.19	.00	-1.08	-.26
Text 2						**Language 2**						**Text 2**					
-.53	.18	-2.89	.00	-.902	-.17	-.42	.21	-2.02	.04	-.83	-.01	-.60	.19	-3.15	.00	-.97	-.22
Text 3						**Text 2**						**Text 3**					
-.73	.19	-3.67	.00	-1.12	-.34	-.95	.19	-5.09	.00	-1.3	-.58	-1.2	.19	-6.07	.00	-1.57	-.80
						Text 3											
						-1.0	.19	-5.23	.00	-1.40	-.64						

[61] Coef = coefficient, Std err = standard error, z and P > z = ordinal logistic regression test statistic and p-value, Min conf and Max conf = confidence interval minimum and maximum

The coefficient values shown in Table 26 do not show causation as such, but give an estimation of the predictive power, i.e. the strength of this specific variable.

The table shows that the results are similar for all parts of the VPT. For active form recognition (part B), compared to the control group (the reference category), glossing (z = 2.01, p = .044) and bold-printing+glossing (z = 2.78, p = .005) lead to significantly different results, which in both cases point in a positive direction. The coefficients for both glossing (.39) and bold-printing+glossing (.54) suggest a strong predictive value of the variable TIE, i.e. reading texts enhanced with glosses or bold-printed and glossed target words aided active recognition significantly. Furthermore, Session 3 turned out to be a significant, positive predictor (z = 4.23, p = .005) with a high coefficient in part B (.87). The test revealed that the factor School featured significantly, as School 2 noted a negative effect (z = -2.63, p = .009, coeff. = -.46). Compared to School 1, students from this school scored significantly lower. Finally, the variable text also contributed to the outcome. Both Text 2 (coeff. = -.53) and Text 3 (coeff. = -.73) yielded significantly lower results than Text 1.

The results look similar for VPT part C, the assumedly most difficult part of the VPT (passive meaning recall). However, unlike in the two other parts, here Language Background turned out to be a significant negative predictor of VPT results (z = -2.02, p = .043), with a coefficient of .42. The bilingual students or those with an L1 other than Danish remembered significantly fewer words than those growing up with Danish only. This matter is addressed in the Discussion. Again, a significant impact was recorded for the factor Session 3 (z = 2.45, p = .000, coeff. = .48). Once more, Text 1 seemed to have been significant easier than the two other texts, as the results for Text 2 (z = -5.09, p = .000, coeff. = -.94) and Text 3 (z = -5.23, p = .000, coeff. = -1.0) were significantly lower.

In VPT part D, which tested passive meaning recognition, the TIEs containing glosses (C2: z = 2.82, p = .005, coeff. = .57, and C3: z = 3.70, p = .000, coeff. = .71) were significantly powerful predictors of the outcome variable, i.e. vocabulary acquisition. As in part C, the students' Language Background turned out to be of significant importance (z = -3.19, p = .001) and carried a negative value (coeff. = -.67). Additionally, the school played a role with respect to how well this variable predicted the VPT outcomes, as the results from students from School 2 show (coeff. = -.38). Finally, it seems that Text 2 (coeff. =-.59) and Text 3 (coeff. = -1.18) contributed significantly to the outcomes of the VPT.

The outcomes of the regression are also revealing regarding how recognition and recall as different facets of vocabulary knowledge were influenced by the different factors. Compared to Session 1[62] the ordinal regression coefficient for active form recognition is large (.88), which means that there is a large significant difference between performance in the first and last

[62] The reference level the other two variants of the variable 'session' were compared to.

session. It seems that the active form recognition task became easier through repeating the procedure three times. The same is true for the passive recall task in part C. The coefficient is, however, smaller (.49). This part of the VPT was difficult and even though going through the procedure three times magnified the scores, the effect is much smaller than for recognition.

These results suggest that the model does capture some of the predictor variables that contribute significantly to the outcome of the VPT. It is striking that across the different vocabulary test parts; the TIEs containing glosses can be linked to significantly better results in the VPT, and that the coefficient increases from bold-printing to bold-printing+glossing. This is in tune with my hypothesis that there may be degrees of enhancement and that less invasive forms of enhancement may aid word learning to a lesser degree than more invasive forms. The fact that Session 3 had a significant impact may allow insights into the role of test familiarity and test fatigue and related to that, into the impact of incidental versus intentional learning. That School was a measurably impacting factor suggests that the students from one of the schools performed at a lower level than those of the other schools. Similarly, the fact that Text 2 and Text 3 both seemed to lead to lower VPT scores as compared to Text 1, confirms my earlier assumption that the texts were not completely equivalent in difficulty. This is discussed in the Discussion. Finally, the participants' language background also contributed significantly to the outcome of the VPT, which shows that many more factors, such as familiarity with reading styles, may be linked to reading (see Discussion).

A final step was to examine model fitting information, to see whether the chosen statistical model actually does improve the ability to predict the outcome as opposed to an empty model, i.e. a model without the predictor variables. The Stata software output provided information about the value of the likelihood ratio chi-square, as the 'goodness of fit' measure and also the value of the Prob>chi-square and the McFadden's pseudo R-square statistics[63]. Likelihood values fall between 0 and 1, a small ratio indicates a better fit of the full model, i.e. the model containing the predictor variables (see Tables Appendix 23, 25 and 27).

In addition to calculating these three model fit statistics, a fitstat measure was run, which provides the results of several other model-fit-tests (Tables Appendix 24, 26 and 28). However, only one of these is reported in Table 23, the Cragg and Uhler's Nagelkerke R-squared test (Cohen et al., 2003; Long & Freese, 2006). According to the Stata Annotated

[63] Prob>chi-square: The probability of obtaining a specific chi-square statistic if the independent variables have no effect on the dependent variable. The p-value, is used to determine if the overall model is statistically significant. McFadden pseudo R-squared: Another goodness of fit measure.

Output, **"**the ratio of the likelihoods reflects the improvement of the full model over the intercept model (the smaller the ratio, the greater the improvement)".[64]
The results of the ordered logistic regression model-fit calculations run for all three parts of the VPT are presented in Table 27.

Table 27. Model Checking Summary - Goodness of Model Fit Indicators

vocabulary post-test part	LR chi2	prob>chi2	McFadden's pseudo R2	Cragg & Uhler's Nagelkerke R2
VocScore B active form recognition	50.34	0.0000	0.0171	0.072
VocScore C Passive meaning recall	116.04	0.0000	0.0405	0.159
VocScore D passive meaning recognition	98.00	0.0000	0.0314	0.136

The values for all four model-fit measures indicate that the model is actually improved by including the chosen independent variables, i.e. these factors do in fact have the ability to predict the outcome variable. The model has an appropriate overall fit and underlines that the data are consistent with the model fitted to it.

To sum up, several of the recorded variables, in particular the TIEs including glosses, the last testing session and two of the texts had in many cases a significant impact on the outcomes in the VPT for active and passive forms of form and meaning recognition and recall. These independent variables interacted strongly with vocabulary acquisition and were able to predict a good proportion of variance in the outcome. Still, even though it was possible with this regression to include many influential variables, this cannot present a comprehensive view of vocabulary acquisition through reading enhanced texts. It is an attempt to describe and analyse the effect of these specific variables on the learning situation.

As mentioned in the introduction to this section, the findings of the ordinal logistic regression also provide answers to the second research question of whether reading a text enhanced in a specific type of TIE supports active or passive forms of either word recognition or word recall in the post-tests significantly better than another form. Therefore, the regression results are examined in light of this question.

An answer to the second research question of which type of TIE would be most beneficial for which type of vocabulary knowledge can be sought by examining the regression coefficients presented in Tables Appendix 23, 25 and 27. For part B of the VPT the highest

[64]Stata Annotated Output. What are pseudo R-squareds? UCLA: Statistical Consulting Group. http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm, accessed on 9/9, 2013.

coefficient was noted for bold-printing+glossing combined (coeff. = .54). Also in the two other parts of the VPT it was working with a text enhanced with bold-printing+glossing that yielded the highest values in the coefficients (part C: coeff. = 1.09; part D: coeff. = .71). In all three cases, the control group was the reference the other enhancement types were compared to. Bold-printing+glossing, it must therefore be assumed, is the most helpful of the three tested TIEs for active and passive word recognition, but also for passive recall. This finding suggests that working with a text in which target words were enhanced through both bold printing and L2 glosses in the margin best supported vocabulary acquisition as assessed in the VPT and thus confirms a trend that emerged in many parts of the analysis.

4.7. Results Of The Delayed Vocabulary Post-Test

Four weeks after the last testing session, the students from six classes were tested again in a delayed VPT session. A total of 107 learners completed this delayed post-test. Unfortunately, only the data of 52 was usable, as only these 52 had participated in all three previous testing sessions.

In the delayed post-test, the participants were asked to complete a VPT comprising of the 30 target words (see Table Appendix 2, Figure Appendix 7 and 8). These data allow insight into whether time is a significantly influential factor and whether the impact of the TIE treatment holds over time.

To investigate these two issues, a censored form of regression, the 'tobit' model, was used (see Equation Appendix 6). This nonlinear model is estimated using iterative maximum likelihood estimation techniques (Scott Long, 1997). According to McDonald & Moffitt (1980:318), it is an appropriate choice, because in it, "it is assumed that the dependent variable has a number of its values clustered at a limiting value, usually zero", which is the case with the VPT scores (scores range between 0-10). In fact, the censored regression model was "designed to estimate linear relationships between variables when there is either left- or right-censoring in the dependent variable" (p. 321).[65]

In the tobit regression, in addition to gender, school, and language background, time was added as a variable. To include this factor, a dummy variable was integrated, where '0' covered the test scores collected in all three testing sessions in the main data collection ('pooled') and '1' comprised of the data collected in the delayed post-test. The main study data served as reference group (to which the delayed post-test data was compared), as did 'male' for the factor 'gender', 'school 1' for 'school' and 'Danish as L1' for 'language background'.

[65] Censoring is conducted when a value occurs outside the range of a measuring instrument. Tobit Analysis. http://www.ats.ucla.edu/stat/stata/dae/tobit.htm accessed 8/9, 2015.

The highest possible score in the delayed post-test was 30, because the 10 target words from each of the three texts were tested in one go in the delayed VPT. To be able to compare the scores from the delayed test with the tests conducted in the main study, the scores for VPT part B from the three main study testing sessions were added up, those for VPT part C and also the points scored in VPT part D, so that the points were scored as x/30[66] [67].

The detailed results of the tobit regression can be found in Tables Appendix 29 – 31. In addition to the regression coefficients, standard errors, z- and significance values, and confidence intervals, the sigma test statistic, the estimated standard error of the regression, is presented in the tables. The Stata output also provided an observation summary containing the number of left-censored, uncensored and right-censored values. The significant results, all concerning the factor Time, are presented in Table 28.

Table 28. Delayed Post-Test Significant Results Based On Tobit Regression, Factor *Time*

Vocabulary post-test part	Coeff.	Std. Err	z	P>|z|	95 % Conf Interval[68]	
					Min	Max
VocScore B active form recognition	-.0838187	.0225256	-3.72	.000	-.1279681	-.0396694
VocScore C passive meaning recall	-.0894492	.0212268	-4.21	.000	-.1310528	-.0478455
VocScore D passive meaning recognition	-.0804378	.0164073	-4.90	.000	-.1125956	-.04828

For all parts of the VPT, time was the only factor which turned out to be a significant predictor of the outcomes of VPT scores when comparing the scores achieved in the main study with those attained in the delayed post-test (VocScoreB: $z(99) = -3.72$, $p<.05$; VocScoreC: $z(99) = -4.21$, $p<.05$; VocScoreD: $z(99) = -4.90$, $p<.05$). The results show that for a one-unit increase in Time, there is a -.08 point decrease in predicted value of the VPT scores in part B, C and part D. However, this proposes that each time this factor explains only a relatively small proportion of the variance (8%). It must be assumed that many factors other than Time played a role here.

For this regression model, goodness of fit measures were also applied. Here, the Wald chi-squared test, which is commonly used for testing in non-linear settings (Scott Long, 1997),

[66] = 3 x 10 possible target word scores.
[67] The equation applied for investigating these variables (vocabulary acquisition as dependent variable and gender, school, language background and time as independent variables), can be found in Equation Appendix 7.
[68] 95% Conf Interval = confidence interval, describes the amount of uncertainty associated with a sample estimate of a population parameter.

and the prob>chi squared test were administered. The results summarized in Table 29 are based on Tables Appendix 29 – 31.

Table 29. Model Checking Summary - Goodness Of Model Fit Indicators For The Tobit Regression.

vocabulary post-test part	Wald chi2	prob>chi2
VocScore B active form recognition	15.31	0.0041
VocScore C passive meaning recall	21.79	0.0002
VocScore D passive meaning recognition	25.57	0.0000

The values indicate that according to both goodness of fit measures the model was a good fit to the observed data for all three parts of the VPT.

To refer back to the two issues investigated in this section, it can be said that in the delayed post-test Time was the only factor that significantly predicted the outcome of the VPT scores. For all three tested forms of vocabulary knowledge it had a negative impact on the test performance. Four weeks after the reading intervention, the participants' performance in the vocabulary tests had dropped significantly. This confirms the influence of the factor *time* and shows that time is a significantly influential negative factor when it comes to the vocabulary knowledge tested in the VPTs.

The post-test was also concerned with whether the effects of the different TIEs measured in the main study were still effective four weeks after the treatment. Unfortunately, it was not possible to do so statistically, because the sample was too small to obtain reliable information, as only 52 of all participants had attended all three testing sessions in the main study and were also present for the delayed test. Secondly, the fact that there are four levels to the independent variable TIE (control group + three TIE-types) caused a zero inflation when dummy variables were added to the tried regression model. Therefore, the question of whether or not the effects of the TIEs observed in the main study, can also be regarded as long-term effects remains unanswered.

4.8. Chapter Summary

In this chapter, the results of the statistical analysis of the VPT data were presented.

The null hypothesis that there is no difference between the different TIEs in their effect on vocabulary learning was rejected. Instead, Hypothesis 2, claiming that TIEs are beneficial for vocabulary acquisition could be generally accepted.

Regarding the research questions under investigation in this thesis, several findings are particularly noteworthy: Across the three testing sessions and across the three tested knowledge types, enhancement forms containing glosses lead to results superior to those yielded from unenhanced texts or those in which target words were bold-printed only. In particular, it turned out that the most effective type of enhancement was the combination of bold-printing and glossing. Moreover, across all sessions and with pooled data, the findings show that it is the ability to passively recall word meaning benefits most from reading enhanced texts (of any type of enhancement).

Furthermore, throughout, the highest scores were achieved in the form-meaning recognition part of the VPT (B), the lowest were recorded for recall (part C). Often, the results from bold-printing are very similar to those of the control group, which shows that bold-printing did not always lead to significantly better results than the unenhanced reading materials. This again underlines the superiority of glossed enhancements for aiding vocabulary acquisition.

The correlation coefficients suggest an overall connection between vocabulary acquisition and TIE, and also a link between the different vocabulary knowledge types, as those participants who performed well in one part of the VPT, also did well in the other parts.

Finally, the ordinal logistic regression revealed that the independent variables (text, session, school, language background, gender) were well chosen as they explain a substantial part of variance in the statistical models; they capture many of the factors that influenced the learning situation. In particular, the significant impact of the factor session shows that the assumed ‘test effect’ has to be taken seriously when interpreting the data, and suggests an incidental – intentional shift between the first and the two following sessions. Above that, a ‘text effect’ was confirmed, which indicates that Text 2 and 3 were more difficult than Text 1.

The following chapter presents the findings yielded from the data collected in the retrospective interviews.

5. QUALITATIVE FINDINGS

5.1. Introduction

It is a widely held view that "quantitative methods alone cannot give a full picture of [… any] phenomenon and that qualitative research designs can make a distinct contribution" (Maggetti et al., 2013:155). The many studies in which different types of data are methodologically combined support this assumption. For example, Bowles (2004:549) in her study comparing computerized and traditional paper-and-pen glosses, came to the conclusion that "a crucial element of language learning may lie not in what external manipulations are made to input but rather in what learners *do* with the input as they interact with it". Her think-aloud protocol data revealed findings regarding the levels of learners' awareness, input and instruction processing and the use of reading strategies that were relevant for understanding vocabulary acquisition through using glosses. These would not have been revealed had she looked at quantitative data only. Therefore, I too included qualitative data in the form of retrospective interviews that sought to expose the interaction between the learners, their vocabulary acquisition[69], and the enhanced reading material. These interviews consisted of two parts. In the first part, the interviewees talked mainly about task and test completion. The second part of the conversation considered the target words in the particular text and how the students dealt with them.

The first section of this chapter (5.2), the 'Theme: Textual Input Enhancements', presents findings related to how well the TIEs were noticed, and how they were used in relation to vocabulary acquisition. In section 5.2.4 the outcomes for the two enhancement types, bold-printing and glossing are juxtaposed and obtrusiveness and the role of the 'feedback' given in the glosses are considered. The chapter continues with presenting data related to the interviewees' ability to recall knowledge about the target words (5.4). In section 5.5, data concerning the strategies which were used to explain and remember the target words and those used for task completion are discussed. The theme presented thereafter (5.6), 'Task, text, and test related issues', discusses statements regarding the effects of these issues.

[69] As in the previous chapters, when I mention vocabulary 'acquisition', it is likely to be an overstatement regarding the type of knowledge that is acquired from one exposure to an unknown word while reading. Still, I apply it for practical reasons, and in most instances I differentiate which type of word knowledge is displayed and why.

Quotes from the interviews are referred to by indicating the pseudonym, followed by the interview session number and the number of the statement within this interview. For example, in 'Runa 2.15', for the 15th turn of the interview conducted with Runa after the second testing session.

5.2. Theme: Textual Input Enhancements

In the reading/testing sessions the TIE were implemented in the form of three reading conditions. As a reminder, these are Condition 1, where the target words were highlighted through bold-printing, Condition 2, where there were glosses provided in the margin to explain the target words, and Condition 3, where the target words were highlighted in bold *and* glossed in the margin.

The 'theme' in this section covers 'topics' that relate to the use of the different TIEs. These include for example the noticing of the enhancements, how they were used and comments related to the impact of the different types of TIE.

5.2.1. Topic: Noticing The Enhancements

Previous research found a direct relationship between how well learners notice input enhancements and success or failure of an intended learning outcome (Bowles, 2004). Merely externally enhanced input, for instance TIE provided by a teacher, can be problematic regarding noticing. It may be that the learners either do not notice the enhancements, or that they notice them only partially, contingent on whether or not they are cognitively ready for them or how much overlap there is between externally and internally generated salience (Han et al., 2008). Given that the underlying purpose of TIEs is to trigger noticing via heightened salience, it seemed relevant to include interview data that related to noticing.

Most interviewees mentioned that they had noticed the TIEs. When Jeppe described how he usually works with TIEs, he pointed out that reading enhanced texts felt 'natural' to him (1.45). He mentioned noticing that some words had been 'enlarged' (1.31). Stina said that she had noticed the highlighted and glossed words already when reading the instructions (3.4). The TIEs were noticed even in Condition 1 where the target words were only bold-printed (Stina 1.10). Students noticed the enhancements even down to single words (Stina 2.56). Given the importance of noticing as an early step in the vocabulary acquisition process, the enhancements can be said to have been successful.

A few statements revealed, however, that the noticing of the TIEs did not necessarily go so far as to differentiate what type of enhancement was used. When reflecting on the three

texts he had read, Jeppe pointed out that he had not been aware of the fact that there had been different enhancement types:

S *"As far as I remember, it is the same, but I guess it isn't, the way you say that … Is it the same?*

I *No. No, it wasn't. But you don't remember what it was like?*

S *No, not really."* (Jeppe 3.5-9)

This and other similar statements (Runa 2.22, Silas 1.46) suggest that learners do not necessarily notice the nature of the different types of textual enhancements, but only that enhancements are provided.

Runa reported that she had seen neither the bold printed target words nor the L2 glosses in Text 2 (2.12). She discovered them only in the interview. Notably, this occurred in her second session, when she knew that she would be tested after reading. This might suggest many things, both related to her as a learner and to the reading material itself: Schmidt (1990) found that failure to benefit from input may arise from a combination of lack of noticing ability on the learner's part and poor input characteristics such as lack of perceptual salience or 'noticeability'. White (1994) found that some participants had not noticed the enhancements. Jourdenais et al. (1995) assumed that this might be because TIEs could be "intrinsically ineffectual" because the participants were accustomed to such enhancements, because they were not ready to attend to the enhanced feature; or because the feature wasn't salient. Hulstijn et al. (1996) suggested that learners sometimes "simply fail to notice the presence of unfamiliar words or believe that they know a word when, in fact, they do not" (p. 327). In my data it might also be that Runa was either rather unimpressed by the fact that she would be tested or that she did not work particularly carefully. Research has highlighted that input needs to be comprehended, before attention can be drawn to form within the input (Han et al., 2008), meaning that effective use of input enhancements is linked to text difficulty and proficiency. Considering that Runa said that she had found the second text more difficult than the first one (2.6), it is likely that she was struggling with comprehension, so that her attentional resources were absorbed by processing meaning so that she could not pay attention to form. This lends support to previous findings that processes for vocabulary acquisition and reading comprehension compete, even at the level of noticing (Ellis, 2004; Han et al., 2008; Rott, 2007).

Furthermore, Sharwood Smith (1991) found that externally induced salience may not necessarily be registered by the learner and even when it is registered, it may not affect the learning mechanisms per se (p. 118). In Runa's case it could also be assumed that she was so anxious that she was unable to register all task features, because later on in the interview she explained:

S "*I get really nervous, when I have to, I know that I something forventes is expected of me."* (Runa 2.30)

By saying this, she seems to confirm this assumption. Her oversight may also have been due to individual preferences. Students work differently and what may be too intrusive for some – which may be the case when working with Condition 3 - may not be sufficient text manipulation for others, depending, among other things, on the strategies used for task completion. This was also found in Bowles' (2004) study, where participants approached the same learning materials using different strategies. This resulted in different levels of noticing of the target words (see also Leow, 2000).

This data shows that whether and how enhancements are noticed is a variable that should be taken into account when interpreting the outcomes of the quantitative data. Without noticing the enhancements or without noticing their specific form at a conscious level, it is unlikely that they are processed further and that the help they are supposed to provide can be used.

It is not always obvious for which purpose TIEs are used. To understand the impact of this lack of clarity on task completion, it is important to see what students use the TIEs for. The next topic covers data that reveals some answers.

5.2.2. Topic: How The Enhancements Were Used

According to O'Donnell (2012:546), "the efficacy of successfully working with textual enhancements is predicated on correct usage […] if readers ignore the glosses then it logically follows that they will neither aid nor hinder comprehension". Therefore, it was important to investigate my data with respect to what was revealed about how the learners used input enhancements. This can give insight into the various different forms 'correct usage' can take in real-life classroom learning situations.

Probably because of the widespread familiarity with TIEs, the usage of enhancements as revealed in the interview data was varied and highly individualised, ranging from ignoring, to meticulously checking every gloss. In general, the use was *need*-based, i.e. linked to text difficulty and learner proficiency (Bland et al., 1990; Ko, 2005; O'Donnell, 2012). Silas' statement in his second interview exemplifies this:

S "*Ehm, usually I just look for the words that I don't know."* (Silas 2.26)

Using the glosses only when needed reflects the approach several interviewees described (e.g. Jeppe 1.35, 3.9, Fie 2.8). This ties in with the motivational *need*-factor of the ILH (Involvement Load Hypothesis, see Literature Review). If the enhanced words are those that are perceived as needed for task completion, the *need*-factor is triggered and TIEs provide the chance to comply with this internal drive.

Furthermore, it emerged that TIEs were used for three major purposes; as vocabulary help, for the reading comprehension questions, and for test preparation.

The glosses were primarily used as instant lexical information during task completion (e.g. Stina 1.18; Jeppe 1.35). Silas applied a strategic approach for using the enhancements as vocabulary help:

S *"[…] I read the words in the margin before I found them here, in fact. So I checked before, then I can keep reading without interrupting by looking in the margin. So, I often looked at the words first […]."* (Silas 2.22)

As he did not want to be interrupted while reading, he read the glosses even before reading the text itself. He used the enhancements like a pre-reading vocabulary list to prepare for uninterrupted reading. He was aware of the fact that using glosses would interrupt the reading flow, so much so, that he adapted his working style. This approach might have had the effect of a pre-reading exercise, i.e. it seems likely that he read the text with a raised awareness of these target words. His statement adds evidence to research about the concept of obtrusiveness (Bell & LeBlanc, 2000; Rott, 2007; Rubinstein, Meyer, & Evans, 2001). His behaviour shows that he did not want to afford the switching cost typically involved with consulting glosses while reading. While this technique probably aids a smoother reading process, the effect on vocabulary acquisition is not clear.

Three of the participants mentioned using the enhancements in connection with comprehension**.** Their statements suggest that comprehension here primarily refers to local comprehension, where specific questions about the text had to be answered, rather than global comprehension. This distinction is relevant as a reading task that focusses on local comprehension potentially triggers a larger *need*-factor to process relevant vocabulary. For this end, TIEs can be used purposefully to guide learners' attention to such vocabulary. Whether and how these target words are subsequently used, however, is a different matter. Stina was one of the interviewees who reported using the TIEs in this context. She said in reaction to having read a text were target words were bold printed:

S *"Some of the words were highlighted.*

I *And did you, did you use that for anything?*

S *Yes, I used it for when I had to answer the [reading comprehension] questions. Ehm, then I could kind of remember where it was, but not completely. But when I found a highlighted word, I like worked up or down from that word. It was easier to navigate around."* (Stina 2.10-12).

(see Fie 2.8, Silas 3.3 for similar responses). Stina displays 'normal' student behaviour; i.e. she checks the task, is interested in getting it done economically, rather than in language learning. This is in line with Bowles (2004), whose think-aloud protocol data revealed that

participants consulted glosses to glean essential meaning from the text and not to learn new vocabulary per se. It is therefore likely that comprehension benefits more from TIEs than vocabulary acquisition, as it is the learners' main concern (see also Ko, 2012). In Stina's case, the highlighted target words helped her to locate certain statements in the text, which was convenient for completing the comprehension questions. Thus, they were used as visual orientation rather than for vocabulary work. For the specific purpose of navigation, the bold-printed target words were used intentionally, but the processing of vocabulary that is likely to have happened alongside with this must be assumed to be incidental.

Due to the prospect of having to complete a vocabulary post-test (VPT) in the second and third testing session, all five interviewees reported that they had used the textual enhancements purposefully in order to prepare for the VPT. This 'test effect' is discussed in section 5.5.3.

The data suggested that the learners used the TIEs in ways that were not limited to help with unknown vocabulary. These different types of use may have had an impact not only on word learning, but also on how well a task was completed overall.

However, many statements confirmed that the students used the TIEs for either reading comprehension or to aid vocabulary acquisition. Data concerning the latter purpose are presented below.

5.2.3. Topic: Textual Enhancements Used To Acquire Word Knowledge

One objective of this study was to find out whether and to what extent consulting TIEs contributed to vocabulary learning. The following section presents interview data that offer some insight.

Jeppe seems to confirm an apparent connection between reading a word's definition in the glosses and acquiring (partial) word knowledge. When asked about the target word 'incarceration' he said:

> S *"[…] because when I read this in the text it was explained in the margin. And because it meant something about sending young people into prison it made my mind, it started off my mind thinking of it, so therefore I now know that word. I might probably not use it, but when I see the word incarceration I know that it means 'something prison'."* (Jeppe 1.99)

While he assumes that he might not yet be able to use it productively, he has acquired receptive knowledge of this word and found a key-word association. Similarly, Silas said that for him, reading a gloss does not necessarily mean that he can use the word (Silas 1.106).

One of Stina's statements also suggests that consulting glosses leads to understanding that mainly reflects receptive knowledge. She said:

S *"And I could kind of remember. Some of them. But, you know, I see a word and then I can read the meaning of it, but afterwards if I have to explain it - that's the tricky part [laughing]."* (Stina 1.20)

She predicts that she would recall the word form, but that it would be more difficult to produce meaning. She describes a phenomenon that even occurs to native speakers, namely that learners can easily comprehend the gist or understand vaguely so that they do not need to interrupt reading to check for meaning. This is different from forcing them to process the word for finer meaning as done in the VPTs. In this sense, the testing situation is a useful scenario for encouraging learners to apply deep processing techniques that might lead to knowledge beyond recognition. Stina's comment seems to confirm that receptive knowledge precedes productive knowledge, and that if TIEs are used with the intention to trigger multifaceted vocabulary acquisition, then greater external effort is needed.

After encountering a word and the gloss only once, word knowledge is likely to be limited (Eckerth & Tavakoli, 2012; Rott, 2007; Tekmen, 2006). Both Stina and Silas point out that using the word actively might yet be beyond their capabilities (see above). However, their statements suggest that consulting the gloss helped establish some memory trace, probably receptive form knowledge, which confirms Schmitt's (2010) finding that at least some knowledge is likely to be registered. As the two interviewees recognize the words and associate them with the correct meaning, their statements show that some sort of intake has been formed through reading the gloss. The three statements discussed in this section so far highlight the structural difference between remembering a word form and actually using a word actively. They also confirm that choosing a vocabulary test which assesses different types of word knowledge was suitable. By differentiating various types of vocabulary knowledge as proposed in Laufer and Goldstein's (2004) taxonomy is meaningful. The Quantitative Findings Chapter presents the detailed VPT results from the recognition parts of the VPT in comparison to the production parts. The Discussion Chapter examines these findings regarding whether and why reading glosses may lead to recognition skills rather than word knowledge skills for active recall.

The statements presented in this section highlight the importance of a clear purpose when providing glosses. This purpose needs to be transparent for the material developer and the learner. Glosses seem to be sufficient only with respect to supporting the reading process and for generating initial word knowledge. However, if the goal is the acquisition of a more multifaceted word knowledge, then the target words should also be practised, for instance in post-reading exercises.

5.2.4. Topic: Bold-Printing Versus Glossing - The Role Of Feedback And Obtrusiveness

Several statements show that the three enhancement types (bold-printing, glossing, bold-printing+glossing) present rather distinct advantages and disadvantages, and may lead to distinct processing of vocabulary knowledge.

Two tendencies became obvious regarding highlighting target words through bold-printing (C1). On the one hand, this was the enhancement type that caused complications. Some interviewees found that bold-printing alone was rather "confusing" (Runa 2.22) and "not helpful" (Silas 1.46). It seems that bold-printing is only perceived as useful when linked to relevant instructions or a clear purpose. Silas explained what he had assumed the purpose of bold printing to be:

S *"I think I should try to find out what they meant. When I saw them here [referring to the bold printed words in the text] instead of just thinking, "Well I don't know what it means.""* (Silas 1.26-28)

It seems that he understood the highlighting of the target words as an invitation not to ignore them, but rather to show their importance, as a reminder that these words should be taken seriously. Without instructions of how and for what purpose the textual enhancements should be used, however, one could not be sure, Silas said (1.44). He had actually asked me what to do with the textual enhancements during the first testing session (Silas 1.16). It seems that enhancing language items without providing clear instructions regarding their purpose might result in guiding learner attention away from them and to an unintended focus on deciphering task requirements (see Silas 1.18; 24; 58). Research found that explicit instructions help focus attention on form and meaning in the input, a prerequisite for subsequent processing (Schmidt, 2001). In this respect, glosses are more straightforward in their use and purpose.

On the other hand, the bold-printed enhancements were deemed useful in several other ways: The highlighted words created a visual structure that was used for practical orientation while reading. One example was what Stina said about using the bold-printed target words for navigating around the text rather than for aiding comprehension when encountering difficult vocabulary (2.10-12, see above). Silas found the highlighted words convenient as they spared him the time to highlight (e.g. underline, draw arrows) difficult words himself; something he typically would have done (Silas 1.56). When asked, Stina not only recalled the target word but also remembered other keywords from the same sentence in which it had appeared (2.52). It seems that bold-printing, which makes target words visually distinct from their context, helps

learners to remember the word form, and to a degree even the context in which they occur, which shows that even such minimally intrusive forms of TIE can be of benefit for learners.

Jeppe pointed out that whether or not it was helpful to bold-print unknown vocabulary, depended on the purpose of the task:

I *"[…] Do you think it is helpful to have words like that highlighted in the text without giving any form of explanation?*

S *Depends on the situation and what you want from us. If you want us to notice the words then yes, highlight them. If you want us to understand the meaning of the text, then no, don't highlight them. […]."* (Jeppe 2.13)

He states that if text comprehension is relevant then it is better not to highlight words, if, however, one wants students to pay attention to unknown words, then highlighting makes sense. He went on to say that wanting both things from students was too much and confusing ("*then you haven't a clue*" 2.13), i.e. the highlighting left him distracted rather than focussed (also Lee, 2007). It can be assumed that he regards the benefit of enhancing texts with Condition 1 as a matter of either successful text comprehension *or* vocabulary acquisition, and that both at the same time are not feasible. This is an important finding, as it puts into question whether enhancing target forms can support vocabulary acquisition and text comprehension simultaneously, without one thriving on cost of the other. This would be in line with research which found that using TIEs to pursue both purposes is of doubtful benefit (e.g. Rott, 2007). Anyone creating enhanced reading materials should take this into account. If both comprehension and vocabulary acquisition are the purpose of the enhancements, then they should be replenished with extra learning support, for instance exercises.

Stina mentioned in her second interview that, while highlighting was fine, glosses would have been better (2.14-15). For her, the advantage of glosses lay in the fact that they made it easier to find the meaning of an unknown word. Fie said that the "small explanations on the side" had been helpful and that she had been relieved that they were there (1.30). By instantly providing the word meaning, the glosses were welcomed as helpful input.

Glosses provide such instant feedback to learners' hypotheses about word meaning. The interviews also show that textual enhancements containing 'feedback' were preferred. This became obvious alongside with the fact that the interviewees were able to pinpoint advantages, but also the disadvantages of using TIEs. Jeppe confirmed that making specific features of a text more salient can be "disruptive" (1.49) and Fie explained how she weighed the positive and negative sides of using TIEs. In reaction to having read a glossed text, she said:

S *"Well, when I, it's like when, when you need, when you go to the point where there are words you don't understand, I stop up and then try to think. But when I just looked up*

> *I saw the word I think, like, ehm a help right from the start instead of just going into yourself, which was good and bad because you didn't get the chance to find out yourself. So, it was good and bad. Because it help[s] but you could maybe have done it yourself. But if you couldn't it was it was fantastic to see it."* (Fie.1.36)

For Fie, then, encountering unknown words in a text makes her stop and think about the word. When glosses are provided, however, she does not feel the need to do so. For her, there are two sides to this. On the one hand, she finds that glosses prevent her from making the effort of trying to find the word's meaning by herself. On the other hand, they might be useful with words where she otherwise would fail to find the meaning herself.

Her statement suggests a potential dilemma: Providing glosses affords convenient instant support for reading, i.e. clear feedback about a word's meaning which, however, cuts short the process of establishing and then testing her language hypotheses, for example regarding this word's meaning. Hence, she does not need to develop strategies for dealing with such vocabulary problems. This is in line with previous studies (Hulstijn & Laufer, 2001; Laufer & Hulstijn, 2001; Hulstijn et al., 1996), where it was argued, that "providing readily accessible word meanings" may have even have a "prohibitive influence on L2 vocabulary learning, as L2 readers are not encouraged to make a mental effort to get the meanings of the words" (Jung, 2016:94). According to Færch, Haastrup and Philipson's (1984) model of hypothesis formation and testing in foreign language acquisition, the interplay between hypothesis formation and feedback is essential for the acquisition process (see Swain's output hypothesis, 1985). Regarding vocabulary learning, Svensson (2012) found a positive effect of feedback on word meaning recall. It can be deduced from Fie's statement that she usually would have tried to infer the word's meaning. When working with reading Condition 1 she would have been forced to think and look for confirmation for her hypothesis. However, she also would not have been given any feedback that would have enabled her to check her hypothesis. However, according to the ILH, the processing effort triggered in glossed conditions is high, as using glosses means de-contextualizing the target word and processing the information in the glosses. This would contradict Fie's perception of what is useful for vocabulary acquisition. In order to understand this dilemma, this issue is taken up in the Discussion.

This 'checking for feedback' effect could be assumed to be aggravated in Condition 3, where bold-printing and glossing were combined. The bold-printed words encourage readers to go to the margin to check the help provided in the glosses. Interestingly, the quantitative scores for Condition 3 sometimes lie lower than for Condition 2 (glossing only). This might indicate that the double-enhancements are too intrusive, in that even readers who are familiar with the word are tempted to check the margin. This checking and interruption of the reading

flow is then not *need*-based, but conducted 'just to make sure' or as a habit. Similar excessive use of glosses was also found by O'Donnell's (2012) and Ko (2012).

This raises the notion of 'obtrusiveness'. As explained in the Literature Review, 'obtrusiveness' refers to the issue of interrupted reading flow caused by consulting glosses. One of Jeppe's statements exemplifies how using glosses is natural, but at the same time comes at a cost ("switching cost", Rubinstein et al., 2001):

> "*I read and as soon, as soon as I reached one of those words, I, you know, like* besetting*, then I would go to the margin, read and then read the two last words from the sentence. So I kind of stopped in the middle of the sentence and checked the word and then I returned. That might not be the best way to do it, but that's what feels best for me. In some way almost natural."* (1:45)

He, as other interviewees (e.g. Runa 2.34, Fie 1.36), interrupts his reading whenever coming across a glossed word, even in the middle of a sentence. He implies that this might be a shortcoming of glossing, but that he is used to working this way. Silas (2.22) even studies all glosses before reading the text in order to not be interrupted. Comparing the VPT scores yielded from un-glossed with those from glossed texts gives insight into whether this obtrusiveness has an effect on vocabulary learning (see Discussion).

The general tendency seems to be that bold printing on the one hand makes readers focus on form, and glosses on the other hand are used like dictionaries. Moreover, glosses are regarded as more helpful than highlighting, not just because they offer feedback, but also because their purpose is transparent (Runa 3.20, Silas 2.22, Fie 3.2). The various types of use of the three enhancement forms may lead to conclusions about different forms of processing. A comparison of the results yielded from the different enhancement forms shows whether there really is evidence for the perceived differences in benefit for vocabulary acquisition (see Discussion).

The findings presented in this section on the one hand highlight the importance of instructions and a clear purpose when enhancing foreign language reading material, and on the other hand show the limits of textual enhancements from the student' perspective.

5.2.5. Topic: Attitudes Towards Working With Textual Input Enhancements

All interviewees expressed a general preference for working with enhanced reading materials as compared to unenhanced texts and specifically favoured enhancements containing glosses. They found having enhancements "very important" (Runa 2.18; Fie 1.30), "very helpful" (Stina 3.6, Jeppe 1.33) and appreciated the instant reading help (Stina 3.18). Fie

specified that even bold printing difficult words made those "more clearly"[70] (Fie 2.10). This is in line with previous research, where subjects' opinions about the value of glosses were positive (e.g. Holley & King, 1971; Jacobs et al., 1994; Ko, 2012; O'Donnell, 2012). In Cheng and Good's (2009) study, for example, 75% of the participants believed that vocabulary glosses could increase English reading comprehension and vocabulary learning. However, Jeppe (1.49) admitted that enhancements could be "disruptive". One statement by Fie (1.36, discussed above), in which she points out that glosses prevent her from making the effort of trying to find the word's meaning herself, could also be interpreted as expressing a critical attitude towards this type of enhancements.

5.3. Theme: Recalling Vocabulary Knowledge in the Interviews

In the second part of the interviews, I discussed the target words with the students. These data give insight into how and whether the TIEs play a role in the learning process[71]. In the following sections, those findings are presented that reveal to what extent vocabulary knowledge was recalled (see section 5.3.1). Section 5.3.2 discusses data that show what hindered the acquisition of recallable vocabulary knowledge.

5.3.1. Topic: Successful Recall Of Vocabulary Knowledge

As discussed before, it is difficult to say what full word knowledge really is. Therefore, 'successful recall of vocabulary knowledge' as discussed in this section denotes partial knowledge, for instance recognition of the word form.

Some interviewees displayed knowledge of both word form and meaning, but in the majority of cases recognition rather than productive skills were displayed. For example, Stina and Jeppe remembered having seen a particular word in the text but were unable to recall its meaning. When asked about the target word 'rigid' in her last interview Stina replied:

S *"No idea.*
I *No idea … Do you think you have seen it before?*
S *Yeah, in the text but …*
I *But you don't know what it means?*
S *No." (Stina 3.39-43, similar Jeppe 2.27, Jeppe 1.63, Jeppe 3.39)*

Seeing the word in the text leaves her aware of the word itself, its form, she says, but she cannot remember the meaning. This is in line with the findings from the quantitative data where

[70] Original mistake.
[71] For reasons explained in the Methodology Chapter, the statements were not considered further in the analysis.

the scores for word recognition were much higher than for recall throughout the three sessions and in reaction to all three forms of enhancement. Some learners are aware of this fact; Jeppe, for instance, said that the recall part of the VPT was more difficult:

S *"[…] Ehm, there is something about explaining that is quite hard. It is easier to understand than to explain."* (Jeppe 1.101, see also Stina 1.54)

This and the previous statement shows that recognizing (word form or meaning) is easier than recalling (word form or meaning).

Context emerged as an aspect influential for successful comprehension and vocabulary acquisition. It turned out that suitable thematic context encouraged learning even beyond the word level. Jeppe and Stina pointed out how helpful they had found the context to remember words. They both demonstrated that they had not only acquired knowledge of particular target words, but also that the target word was actually part of a complete thematic chunk that was remembered. Stina remembered visually where the word had been located in the text, but also the surrounding information when she discussed the target word 'sphere':

S *"[…] its top of maybe the first page […] I think it has ehm, in a sentence with Obama"* (Stina 2.52-54, similarly Jeppe 2.31).

She recalled the immediate context, the textual layout and the surrounding sentence in which the word had occurred. This is in line with research into the potential importance of contextual clues for vocabulary acquisition (Hulstijn et al., 1996; O'Donnell, 2012; Svensson, 2012).

The findings presented above are based on the analysis of the interview data with respect to what kind of word knowledge was acquired. Encountering an unknown word while reading, it seems, offers the possibility to acquire at least partial knowledge of this word and in some cases even knowledge of several dimensions of the word, like form and meaning. This is especially true if the text is enhanced to highlight and explain these low-frequency words. These findings confirm the results from the quantitative data analysis, in that Condition 3 most often led to scores significantly higher than those from the other enhancement types. Some productive knowledge of the target words was displayed. However, the knowledge reported in these statements presents initial knowledge, which was assessed immediately after exposure to the target words.

Almost all of the examples of at least partial word knowledge recall were found in the second or third session. This suggests that the interviewees had become more familiar with the reading and testing procedure, had probably approached the target words intentionally, and that this had aided acquisition. Thus, it confirms that splitting up the VPT data into those collected after the first, second and third session had been sensible, because it allowed for a comparison between incidental (Session 1) and intentional learning (sessions 2 and 3).

5.3.2. Topic: Difficulties With Recalling Vocabulary Knowledge

Previous TIE-studies identified several sources for difficulties. O'Donnell (2012) found that readers misread glosses or had problems with matching the target word with the appropriate corresponding gloss. Lee (2007) saw problems rather in the required amount of decoding. In my study, the interviewees were often not able to recall anything about the encountered target words, and word meaning could not be retrieved, nor did the word look familiar (e.g. Jeppe 2.25, Runa 3.52, Silas 1.112, Fie 2.18, Stina 2.36). This confirms that the target words were largely unknown prior to testing and that encountering them once is not likely to leave memory traces that lead to recallable knowledge. Some interviewees, however, were convinced that they remembered word form or meaning, but they were mistaken. Several aspects emerged as important here. One was that interviewees drew incorrect parallels to similar sounding, seemingly familiar higher-frequency words. Examples are the target word 'staple', which was confused with 'stable' (Stina 1.54, Jeppe 2.35), or the target word 'poised' which was mixed-up with 'poison' (Stina 2.26-28), or 'suffrage' which was wrongly analogized with 'suffering' (Stina 2.46). This was commonly found in the quantitative data, across all participants and enhancement types, and confirms research in which learners overestimated their knowledge because they mixed up word forms with similar looking words, so called 'synforms'[72] (Laufer, 1989; 1997; Laufer & Yano, 2001).

Another problem with recall was revealed when Jeppe discussed the target word 'ubiquitous'. When talking about it, he pronounced the word incorrectly and mentioned how problematic it was for him to remember the word's meaning. He points out that these two aspects might be linked, that his problems with remembering the word might be connected to his inability to pronounce it correctly:

S *"… that would be ubiquitous [pronouncing it wrongly].*

I *Ubiquitous. It is such a weird word, isn't it?*

S *Yeah, and that might be the reason for me not to know it."* (Jeppe 3.45, similarly Runa 3.50).

He supposes that a word that is unusual in its pronunciation may be particularly challenging to learn. This may be because certain basic facets of a word need to be in place in order to enable learners to approach acquisition successfully. Research has found that this really is the case as word form is strongly linked to the way a word is pronounced (Erler, 2007; Laufer, 2012).

[72] 'Synforms' are similar in form but not in meaning to other words (Laufer, 1997a).

It may be that the participants had acquired knowledge, for instance pre-consciously noticed the word's phonological structure, which could not be formulated in the interviews. For this to happen, deeper processing would have been needed. Accordingly, the high number of incidents where interviewees failed to recall any part of the target words is not surprising. Expressible word knowledge can hardly be expected after only one encounter and without any means taken to support word learning further, even for intentional learning. As vocabulary learning largely is a subconscious process (Ellis, 1994), the interview data can only give a superficial insight, as merely conscious processes are described here. However, the statements discussed in this section might prove insightful when looked at together with findings from the quantitative data in the Discussion Chapter.

5.4. Theme: Vocabulary Strategies

Bowles (2004:549) investigated learner behaviour and found that "even given identical materials, all students will not process them alike". I wanted to investigate which strategies learners used for reading and vocabulary acquisition, so that approaches to TIEs might become apparent. This section explores strategies used to explain and remember the target words (section 5.4.1) and strategies used for task completion (section 5.4.2).

Different types of strategies were applied to deal with new and difficult vocabulary. For example, the interviewees reported that they had used repetition or guessing while reading (Silas 1.90. Runa 1.115, Jeppe 3.9). Additionally, four of the five interview participants used synonyms as a way of explaining the meaning of a target word in the interviews (e.g. Stina 1.34, Silas 1.74, Runa 1.62). Another strategy was to link the target words to emotional associations in order to explain them (Runa 1.117, Jeppe 1.85, Stina 1.44, Stina 1.56). Also the use of L1 related strategies or related to a foreign language other than English was mentioned (e.g. Fie 2.34), where the interviewees compared target words to L1 cognates (Silas 3.30, Jeppe 2.15) or tried to translate them (Runa 1.56). However, while previous research found the L1 to have an important impact on L2 strategy use (Folse, 2012; Nation, 1982), it did not, except for some few statements, appear much in my interview data. I only considered the strategies relevant to understand the VPT outcomes better, in particular those, which interact with the TIEs.

In many cases, the learners described using a combination of different strategies to solve a vocabulary problem. In this section the strategies used by the students are considered in different categories for the purpose of illustrating specific points, unless combined use of strategies was relevant in order to understand what the interviewees did and for what purpose.

I categorized the strategies in my interview data based on Schmitt's (1997) strategy taxonomy. Most of the used Vocabulary Learning Strategies (VLS) fall into what Schmitt categorized as "memory" strategies for "consolidating a word once it has been encountered" (e.g. connect word to a personal experience, associate the words with its coordinates, connect the word to its synonyms and antonyms; p. 207). However, cognitive and metacognitive strategies were also used for consolidation, and "determination" strategies (e.g. analyse part of speech, analyse affixed and roots) were applied to "discover" word meaning (Schmitt, 1997:207f). It seems that sometimes my participants used what Schmitt categorizes as "consolidation" strategy, to "discover" a word's meaning. As the spectrum of strategies in my data also comprises task completion-, test completion-, and reading-strategies, I added those. Due to specific circumstances of the testing situation, I further distinguish between strategies used for remembering and explaining vocabulary. Thus, Schmitt's taxonomy covers and structures most of the VLS apparent in my data, but it had to be adapted.

5.4.1. Topic: Strategies Used To Explain Or Remember Vocabulary

This section gives insight into which strategies were used to explain and remember the target words in the second half of the interviews. Strategies elicited from the statements about text reading in the first part of the interviews are also included. Even though the two activities of remembering and explaining are distinct mental processes, they are considered within one category in my analysis. This is because the interview data did not allow for a clear differentiation between those two activities.

Ignoring unknown words and different forms of lexical inferencing were found to be the most commonly used strategies. In this aspect, my data are similar to data investigated in previous studies (Fraser, 1999; Laufer & Sim, 1985; Paribakht & Wesche, 1999). My interview data contained numerous statements, which revealed the use of a wide range of other VLS. Most strikingly, the interviewees drew parallels to similar looking/sounding words, used contextual clues in particular for comprehension and often in combination with related cognitive strategies such as guessing or skipping unknown words (e.g. Jeppe 3.7-9). Such statements confirmed that a text's topic matter and genre influenced the use of strategies. The interviewees also reported that they had sounded out the unknown words or split them up in smaller seemingly familiar parts (Fie 1.44, Stina 2.18, Silas 3.11; Jeppe 2.27).

Most relevant for the questions under investigation, three interviewees talked about the fact that they had created their own glosses and typographical links to unknown words. As explained in the Literature Review, whether or not enhancements are externally or internally imposed can have a significant influence on how effectively they are used. Previous research suggested that the difference between internally and externally initiated enhancements is an important one (Han et al. 2008; Ko, 2012; Sharwood Smith, 1991), as external enhancements

may not overlap with the needs of the learners. For my participants, creating their own enhancements was one strategy used in the attempt to fill vocabulary knowledge gaps while reading. Runa, Jeppe and Silas reported working this way. In his second interview, Jeppe said the following when describing how he dealt with unknown words:

S *"[…] sometimes I would underline the word I don't know and later I would have given it a meaning. […]"* (2.4)

He underlines the words and usually later allocates a meaning-definition. It seems that underlining helped him to raise the salience of the word so that he later on, with a greater knowledge of the text's topic was able to assign meaning. Runa described using a similar strategy (2.34). Their method is interesting because, it means that they by choice interrupt the reading process to focus on vocabulary and that they believe this strategy to be useful. If this is a habit, it suggests that they have integrated an intentional attitude to vocabulary acquisition, independent of task instructions. This demonstrates that Sharwood-Smith's (1993) internal/external distinction is relevant, and also that the incidental/intentional distinction is difficult. A similar, yet slightly more sophisticated way of creating self-made enhancements was described by Silas. He reported that he usually creates his own glosses by writing in the margin:

S *"and then [I] just make an arrow or something to tell myself what it means"* (Silas 1.54).

This method underlines that some sort of visual support - here drawing an arrow - is regarded as useful. This is important as this student said earlier on in the interview that bold printing alone was confusing and not helpful. Later on, he modified this statement by explaining that either bold printing or drawing an arrow was helpful, but that either one of them would be needed in order to be advantageous:

S *"[…] when they are bold print I don't need to make arrows and that can help me a bit, then I just write it here [i.e. in the margin] and then I know that it's the bold print words."* (Silas 1.56)

This highlights the important role that typographical enhancements have as visual support while reading. Typographical enhancements like bold-printing target words can create a 'practical' advantage for the learner: encountering the difficult words in bold print makes it easier to remember which of the words might be important and therefore adding further "typographic cues" like for example drawing an arrow may not be essential anymore (Jourdenais et al., 1995). It is merely necessary to add a mark in the margin.

In the interviews, I not only enquired about the target words, but was also interested in which strategies the learners had applied regarding reading comprehension and vocabulary acquisition, and whether any of these were linked to the TIEs. The findings are presented below.

5.4.2. Topic: Task Completion Strategies

Rereading parts of the text and skipping/ignoring difficult parts of text were the two reading strategies that were most commonly mentioned. In many instances the use of these strategies was described in connection with the TIEs, for instance when creating enhancements, or when reading the glosses meticulously. Schmitt's (1997) taxonomy categorizes rereading ("repetition") as a cognitive and skipping a new word as a metacognitive consolidation strategy. In my interviews, reading strategies used for comprehension could not be separated from reading for vocabulary learning.

Most interview statements concerning reading strategies conveyed that the participants regarded global comprehension as sufficient for completing the tasks. This is despite the fact that part of the reading task was answering comprehension questions, which focussed on understanding at sentence or even word level and would have required local comprehension. This shows that strategy use not only depends on the task, but also on learner preferences and learning habits. Thus, the interview data confirms previous research (Gu, 2012; Lawson & Hogben, 1996; Nassaji, 2004). One of Jeppe's statements exemplifies this approach:

S *"[…] I was trying to understand the text ehm, like not specific parts of the text, but the whole text as a unit, ehm, so that's kind of what I focussed on."* (Jeppe 1.23)

The main goal for him was global text comprehension, and when he focussed on unknown vocabulary, it was with the purpose of comprehension, not to learn these words. According to the interviewees, these two strategies, ignoring and reading for comprehension, interact in two ways. On the one hand, difficult vocabulary is ignored because global comprehension can be achieved without it ("*usually if there if there is a word I don't know I just read further and then usually, ehm, it ehm makes, makes sense*." (Fie 1.6, see also Silas 1.48). If, on the other hand, knowledge of difficult vocabulary cannot be gained, then it might be possible to compensate for this lack of specific knowledge with global comprehension:

> "*I would move on and then hope I would understand it, the sentence, when I had read it and if I don't even, maybe I just go on and hope that I get you know the big point out of the story and if I have a dictionary I would look at it, I think."*[73] (Jeppe 1.51, see also Fie 2.8).

Therefore, if target words are to be learned, they have to be meaningful and relevant for task completion – a finding that confirms previous research (Laufer & Hulstijn, 2001; Shook, 1999; VanPatten, 2002) – lest they may be ignored.

[73] Please note that the same statement has already been discussed in the context of task approach (see section 3.2).

When the interviewees regarded local comprehension as one goal of the task, they used TIE-related strategies. Silas linked these with strategic rereading.

> S "*I read the first page. I read all of it and then I looked at the questions quickly and then I started reading. And I read it from start to the end. So, I read all of it. Then I read some of the lines twice with the marked words and I also read the text in the margin to understand the words. And when I had done that I went back to the first questions, answered them and tried to look a bit in the text to ensure that my answers were correct. And then afterwards I went to the last page with questions* […]*"* (Silas 3.3)

He describes a task approach, where strategies (reading for gist, reading intensely, focussing on glosses, rereading comprehension questions, rereading relevant parts of the text, going back and forth) and purposes (overall task completion, global and local comprehension, receiving vocabulary help, answering questions, checking answers) overlap. Several other interviewees mentioned rereading as one of the strategies that related to both global and local comprehension, and also vocabulary acquisition (e.g. Fie 1.12, Jeppe 2.11, Runa 1.58). Thus, rereading was applied to both difficult words and to larger chunks of text. Jeppe describes rereading as prompted by the TIEs:

> *"I stopped more or less when I saw those words and read the sentence. Then I read it again and maybe one more time and trying to figure out what it meant."* (2.11)

When encountering enhanced words he reread full sentences ("section repeating", Nassaji, 2004:119) several times to approach the meaning of the target words. This links rereading in the context of TIE-use closely to the notion of obtrusiveness: The TIEs prompted him to interrupt the reading flow purposefully to apply a strategy that would aid meaning processing. I interpret this as a sign of intentional learning. Runa said that rereading is a typical way of compensating for failed comprehension, for not "thinking enough" while reading (1.58). Stina describes rereading as a strategy applied for vocabulary comprehension when she says: "*of course there were some words that I just had to read twice but, you know, it was fine*" (1.10, similar 2.2).

The reported strategies describe behaviour that sometimes directly relates to the use of TIEs and sometimes gives insight into more generic ways of handling gaps in vocabulary knowledge while reading. The Discussion connects these to the outcomes of the quantitative data analysis. Below, I present data that show how task, text, and test had an impact on the outcomes of the VPT.

5.5. Theme: Task, Text, And Test Related Issues

The nature of the reading task, the chosen texts, and the subsequent test clearly emerged as themes in the interviews, as the interviewees discussed issues related to these aspects many times. Related data are presented in the following sections.

The interviewees mentioned the reading task with respect to several issues that emerged as the topics 'task difficulty' (section 5.5.1), 'task approach' (section 5.5.2), and 'text effect' (section 5.5.3). These are considered here as they shed light on which task- and learner-inherent issues had an impact on the outcomes of the VPTs. Several statements related to test completion and revealed that the test had influenced the participants' task approach. This consequently suggested that 'test effect' (see section 5.5.4) and 'test difficulty' (see section 5.5.5) had to be considered to understand the outcomes of my study and to evaluate my methodology.

5.5.1. Topic: Task Difficulty

As intended in the task design, the target words were perceived as 'difficult' (e.g. Stina 2.2, Runa 1.44, Fie 2.4, Jeppe 3.1, Silas 1.34). However, sometimes the task structure and the instructions were also regarded as difficult, as was shown by unelicited statements from three interviewees. Two interviewees said that the source of difficulty lay in the structure of how the task was presented. Jeppe and Runa were "confused" about the order of the task. As the comprehension questions were placed before the text, the task was structured differently from what they were used to (Jeppe 1.25, Runa 1.24)[74]. Consequently, Runa found it difficult to get started. She said:

> S *"At first I am, I was not sure if I just could start and reading. But then I read the lines [the reading comprehension questions] and then I was a little bit confused if we got the article […]."* (Runa 1.24)

This change in the order of the task materials made her wonder whether and how to start and whether there was a text at all. Runa's statement raises the question of whether this unfamiliar task structure may have impacted on her task completion. If learners have to think too much about how to approach a task they might get distracted. It suggests that feeling accustomed to task instructions, i.e. 'task familiarity', is a concept worth considering in learning and testing situations. This is confirmed by Révész' and Brunfaut's (2013) research. They investigated whether learners' perceptions of text and task difficulty related to the actual difficulty of the tasks and found that the two facets correlated strongly.

[74] This order had been implemented according to pedagogical considerations and in agreement with the teachers' recommendations.

The problems described by Silas were also relevant for understanding TIE-related learner difficulties. He could not see the point in highlighting words without instructions. The given instructions merely pointed out that the enhancements were there to help with the reading. When he was asked to elaborate on this problem, he explained:

S *"[…] do you think it helps you to think about that you have them [the target words] there in bold print?*

S *Yeah, it could but I didn't really think of it, if I had been told that I should take notes of these words before, then I could use it, but I didn't really use it."* (Silas 1.58)

His reaction suggests that only bold printing target words may or may not raise a reader's level of awareness of these words, and if at all then not necessarily for the good of the learner or better task completion. Rather, it may merely cause confusion. It seems that for noticing or (deeper) processing to be triggered, only specific forms of enhancement are useful, most likely forms familiar to the learners. Otherwise clear instructions are needed.

Task difficulties, caused for example by an unfamiliar task order, may cause confusion and take away attention from language learning processes as learners are likely to have only limited attentional resources.

5.5.2. Topic: Task Approach

The statements explored here show how the learners worked with the reading tasks and how much time and effort they found necessary to invest into completing the task. It was striking that several participants seemed to have approached the tasks 'economically'. They completed the task by investing as little time or effort as necessary. Runa critically reflected on her economical task approach with respect to dealing with difficult vocabulary after having read an unenhanced text:

S *"[…] altså [well], when I read something in English, I think in English, but I don't, ehm oversaette?*

I *Translate?*

S *Yeah, I don't translate the Danish when I read it. So, sometimes you just read it and don't think that you understand it. So sometimes it's … you should just stop and think 'Do I understand what I am just reading?' Then I have to read it again."* (Runa 1.56-58)

Her statement suggests that she knows that reading by continuing with a text without ensuring sufficient comprehension may lead to difficulties. She seems aware of potentially helpful strategies (translating, rereading, reflecting) and of the fact that superficial reading may create extra work. Her statement is interesting as it underlines the importance of the *need*-factor (see section 2.3.5 on the Involvement Load Hypothesis). A task, which asks learners to focus on

local comprehension and thus on specific vocabulary as integral requirement may compel learners to reflect and process words. As she says, when not forced to deal with knowledge gaps, she continues reading 'economically', i.e. to get through the text, while not paying much attention to content, with 'reading again' as a possible solution. Her predicament underlines the usefulness and convenience of working with glosses, through the instant feedback that is provided. Thus, TIEs seem likely to influence learner motivation. Stina (1.14) made a similar statement, which showed that task completion rather than reading, and least of all language acquisition, was on the agenda. According to these statements, the students are reading to 'get through' the task rather than for understanding or pleasure and seemingly least of all in order to learn words.

The interviews conducted after the first reading/testing session show an incidental vocabulary learning. When the interviewees described their working procedure, there was no indication that they had purposefully focussed on the unknown words. Rather, they had been concerned with comprehension and seemed genuinely surprised by the VPT.

Difficulties related to task approach might have arisen because learners did not find it necessary to change their typical working style. Fie started with reading the instructions on the first page, but skipped the reading comprehension questions and went straight to the text, where she first read the headlines only (1.8). She explained that she just wanted to start quickly ("*So it is just going to it immediately.*" 1.14), rather than consulting the comprehension questions. The same is true for Silas, who read the instructions, but ignored them anyway and continued working the way he always works (Silas 1.20, also Jeppe 1.51). These students must have assumed that they knew what this task was about, which skills were required, and that working according to habit would be suitable. They trust their instincts rather than following instructions. It appears that while placing the reading comprehension questions before the text breaks the familiar working patters of some students, for others it does not seem worth the effort to change anything. They would rather continue their usual economic working style.

Working economically, it seems, is closely related to working habits. Learners follow their habits to a large extent and breaking these habits probably requires more than just rearranging the structure of a task, for example clear instructions to work differently. Adhering to their familiar habits may result in difficulties in task completion.

5.5.3. Topic: Text Effect

Text-inherent features contributed to comprehension as well as to test performance. The interviews suggest that text-related aspects like topic familiarity and text genre were relevant, as several interviewees repeatedly mentioned them (e.g. Runa 2.5). These aspects are

interesting, because the quantitative analysis also shows results that are somewhat different for the three texts across students from all participating classes. Text genre and text difficulty might have influenced this outcome.

One text-related issue brought up by all five interviewees was topic familiarity. When Stina mentioned that she had found Text 2 more difficult than Text 1, she explained that this was not only because of the many difficult words in this text, but also because she was "not into politics", the topic Text 2 deals with (Stina 2.4; see Fie 2.4, Runa 2.6, Silas 2.2 for similar responses). Jeppe found reading the text in his second testing session easy. He, too, explained that this had to do with the topic:

S *"It was quite easy, I think. I have been reading things in a similar subject and therefore I understand quite a lot words and I knew something about the genre and all that."* (Jeppe 2.2, similar Runa 1.40)

He was familiar with the subject, the genre, and assumed therefore that he knew many of the used words. His statement confirms the importance of topic familiarity for aspects like vocabulary recognition and successful reading and he points to text genre as one aspect that made it easier to understand the text. This positive impact of genre and topic familiarity on reading comprehension has also been found in related studies (Alptekin, 2011; Lee, 2006; 2007; Shafizadeh & Sajedi, 2013, Overstreet, 1998; Waring & Takaki, 2003).

Both topic- and genre familiarity were mentioned as aspects relevant in connection to vocabulary acquisition. The interviewees confirmed that both aspects, in their view, had an impact on the familiarity of the vocabulary and therefore how they dealt with the task.

5.5.4. Topic: Test Effect

The interview statements often evolved around the VPTs, which suggests that the learners were rather preoccupied by the test. This occurred despite the fact that they had been assured that their test-performance would not be discussed with their teachers and that their data would be treated anonymously. Their raised awareness must be assumed to have had an effect on task approach and test completion. Therefore, it was important to consider this aspect here. It turned out that most of the statements in this context were related to two topics, 'test effect' and 'test difficulty'.

All interviewees (e.g. Jeppe 2.5, Silas 3.3, Stina 3.5, Fie 2.5., Runa 3.14) disclosed in the second and third interview session that they had used the textual enhancements increasingly intentionally in anticipation of the VPT. They said that they even deviated from their usual way

of working with input enhancements and adapted the way they used the text adaptations for that purpose. I called this phenomenon 'test effect'.[75]

Some of the statements were already discussed above: For example Stina's, who mentioned reading all the glosses before even starting to read the text itself and then reading them again repeatedly while completing the task, all because she knew that the words would be tested. She wanted to make sure that she was well prepared for the test (Stina 3.18; 3.46, similarly Runa 3.14). Silas, too, described making an extra effort to learn the words in the margin. What he said suggests that the upcoming test had a strong effect on his task approach:

S *"Well, I knew the idea of the whole test, so I tried to memorize the words and tried to find my way of remembering it. And reading the sentence, the line where it was used a couple of times and I tried to use some of the words with myself to remember the words better and say them out loud."* (Silas 3.11)

He believes to have a good idea of the purpose of the test. In his previous interview, he had already mentioned that he assumed the test to be "about learning new words through the reading" (Silas 2.24). Accordingly, he attempted to remember the highlighted words especially carefully. Several aspects in his statement are noteworthy. One is that he was so well aware of the purpose of the test and knew what was relevant for performing well in it. This awareness existed despite the fact that the participants were informed about the purpose of this study in only very vague terms[76]. However, as these students obviously adapted their strategies, there seems to be a clear test effect. Silas mentioned having used several strategies that he seems to have adapted especially to the circumstances of the reading/testing session. This assumption was confirmed in a statement in the second interview when he pointed out that he usually only consulted the glosses for words unknown to him. For the reading sessions in my study, however, he diligently read all the glosses (Silas 2.26).

This suggests that using the same procedure to test the students three consecutive times led to test familiarity that influenced participant behaviour. The incidental vocabulary learning had shifted to intentional learning. This test effect may have led to data that might not reflect natural learning conditions, but somewhat artificial, test-driven circumstances.

While it is difficult to measure the strength of this effect on participants who were not interviewed[77], it must be assumed that at least some of them completed the tasks with an

[75] This term has to be differentiated from 'testing effect' as it is used in learning psychology, where it is applied to refer to how test taking, in particular memory tests, enhances the long-term retention of the information encountered in this test (Carrier & Pashler, 1992), which is different from what I am describing.

[76] In order to not distort the test results and not to steer the participants' behaviour.

[77] See, however, the measures taken when analysing the quantitative data – splitting up the data in sessions and considering them as separate stages of data collection as well as in combination.

awareness of the purpose of the study or at least with a growing familiarity of the test procedure and possibly with some degree of test fatigue.

The interview statements seem a strong indicator of the fact that repeated measures testing methodology has its shortcomings: in some instances, the test effect seems to direct the task approach to a degree that it unquestionably should be taken into consideration when using a repeated measures testing approach. Thus, it is relevant when considering the test results here, but also raises questions about the test validity of other studies with a similar test set-up (e.g. Brown et al., 2008; Newton, 2013; Vidal, 2011). This problem was considered in the Discussion Chapter.

5.5.5. Topic: Test Difficulty

This topic deals with statements related to what made the test easy or difficult.

One aspect causing difficulties was test form. Part B and D of the VPT were multiple-choice questions. These led some students to approaching the test logically rather than through using their word knowledge. One of Jeppe's statements exemplifies this. He described why he found the test type easy to complete:

S *"And actually there with [the target word] 'notion', I said [in part A of the vocabulary post-test] that I didn't know what it meant. But here [in part B] I just said, okay, it is the only word out of these four options that is from the text. So, probably, it is that word. It was a guess […]."* (Jeppe 2.47)

He explains that he assumed one word that seemed unfamiliar in part A later on to be the correct target word because it was the only one he remembered from the text and not because he could match it with the suggested definition. He must have acquired some word knowledge, as he seems to recall the word form. Fie also said that multiple-choice seemed easy, because it gives a choice and because one could simply exclude the least likely other options:

S *"When it is like this [part D of the vocabulary post-test], it is much easier, because it is not that, it is not that and it is not that ..."* (Fie 2.40, see also Jeppe 2.4).

Silas used the same technique of excluding the least likely options, but did not succeed in finding the correct word (Silas 2, 70-72). As three out of the five interviewees mentioned this strategy, it seems to represent a rather common approach. This raises issues of test validity that was taken into account when analysing the explanatory power of the quantitative data in the Discussion Chapter.

5.6. Chapter Summary

This chapter reviewed themes and topics from the interview data, which were relevant for understanding the effects of TIEs on vocabulary learning. The interviews revealed differences

between the three enhancement types with respect to how they were noticed, perceived, how they were used, and for which purpose. Despite their many potential benefits, it was shown that textual enhancements can also have a negative impact on language learning situations. Often, due to practical reasons, there is a disparity between how textual enhancements actually work versus how they are intended to work. For example, bold-printing was often perceived as confusing, while glosses were mostly preferred, but both enhancement types were used for purposes beyond comprehension or vocabulary acquisition. The data showed that the participants were aware of the advantages and disadvantages of the different enhancement types, for instance regarding their level of obtrusiveness and the effect of 'feedback' in the glosses, and that enhancements are not necessarily helpful. These finding reveal dilemmas, which could explain some of the unclear findings previous research produced. However, they do also confirm several assumptions made by previous researchers, for example that Sharwood Smith's (1993) original distinction of internal versus external enhancements is substantiated; that using TIEs with the intention of furthering both reading comprehension and vocabulary acquisition might not succeed (Bell & LeBlanc 2000; Jung, 2016), and that the conceptual *need*-factor (as claimed in the Involvement Load Hypothesis, Laufer & Hulstijn, 2001) is reflected in real-life learning situations.

The data also revealed the principles, which underpinned learner behaviour in relation to TIEs. It emerged that learners follow a habit-driven and economical working style, sometimes even despite task instructions. Still, the statements made it obvious that clear instructions were wanted for all forms of enhancements. The fact that learners reported the common use of self-made enhancements shows how natural it is for them to work with enhancements.

Furthermore, the interviews confirmed that splitting up the data for the analysis into those collected after Session 1 and after sessions 2 and 3 was substantiated, as the learners' learning clearly shifted from incidental to intentional after Session 1. With this, a test effect became increasingly obvious.

All these findings show that adding a qualitative dimension to my data was valuable as the interview statements contribute to a better understanding of the issues investigated in the research questions. In the following chapter I discuss how these findings inform the outcomes of the quantitative data analysis.

6. DISCUSSION

6.1. Introduction

In this chapter, results of the quantitative and qualitative data analysis are discussed in relation to the relevant research literature. I will exclusively discuss those findings where a mixed method approach lead to insights that go beyond what has been discussed in the previous two chapters. The structure of this chapter is based on the research questions investigated in my study.

In this chapter, the term vocabulary *acquisition* is again used cautiously to refer to the initial stages of vocabulary knowledge that are likely to be activated when encountering an unknown word once during reading (see Introduction for details).

6.2. Research Question 1: "What are the immediate and long-term effects of different forms of textual input enhancement on incidental and intentional learning of EFL-vocabulary from reading tasks?"

The already discussed findings show the different effects of typographical enhancements and lexical enhancements containing glosses on vocabulary acquisition. They also relate to the concept of 'obtrusiveness', and shed light on how using enhancements is linked to incidental or intentional word learning from reading.

The outcomes of the statistical analysis regarding the effects of textual input enhancements (TIEs) on different types of vocabulary knowledge look distinctly different for all reading/testing three sessions (compare Tables 11, 14, and 17). Drawing generalizable conclusions is therefore difficult. In several cases, however, the interview data was helpful for interpreting these results.

Outcomes gained from reading texts containing bold-printed target words (Condition 1, C1) were different from those containing glossed enhancements (Condition 2, C2 and Condition 3, C3). A potential advantage of bold-printing is that it does not interrupt the reading flow and readers can read on so that noticing occurrs without overt interruption. Bold-printing target words is a deliberate attempt to raise the level of noticing that is commonly used by teachers and that is assumed to guide learners' focus on word form, a vital ingredient for further

processing (LaBrozzi, 2016; Leow, 2000; Robinson, 2005; VanPatten, 1990; 2002). Previous research findings suggest that, even though the effect of such typographic enhancements is usually small, it can be useful for vocabulary learning, for instance to show that certain vocabulary belongs together (e.g. phrasal verbs, collocations, Bishop, 2004; Boers et al., 2016). However, in my study, both in the statistical analysis and in the interviews, bold-printing yielded inconclusive results: The outcomes were often similar to those of the control group, in that significant group differences with other conditions were rarely found. In many cases, the effect of bold-printing was not profound enough to affect initial vocabulary learning significantly differently from reading an unenhanced text. It was only in in the third session that a significant group difference was found between the control- (C0) and the bold-printing group (C1), and only regarding passive meaning recall (test part C). Except for this case, it made no significant difference whether a learner read an unenhanced text or a text with bold-printed target words.

As bold-printing raises salience merely typographically, it is likely that learners would notice word form, rather than meaning (Boers et al., 2016). Generally, in my study the validity of this assumption was supported by the way the interviewees explained the target words in the interviews. They primarily resorted to form-related descriptions after having worked with Condition 1 and used more meaning-related explanations after having worked with enhancement types containing glosses (Qualitative Findings). Stina made this particularly explicit when she remembered that a specific word had been highlighted in the text (= form), even without knowing its meaning (2.56). This confirms the results of research where typographical types of enhancement led to a specific focus on word form (LaBrozzi, 2016) or the form of a grammatical phenomenon (e.g. English plural markers in Simard, 2009). Accordingly, when working with bold-printing as enhancement, the participants found the recall task more demanding than the recognition task[78]. This provides a potential explanation as to why the significant group difference was found in the recall rather than recognition part of the vocabulary post-test (VPT). In the recall task, even a 'weak' enhancement form like bold-printing was helpful and made a significant difference as compared to an unenhanced text, even in the third testing session, when some participants were increasingly aware of the test procedure and the purpose of the VPT.

According to my data, often bold printing did not enhance the target words strongly enough to lead to outcomes significant from reading unenhanced texts. Similarly, LaBrozzi (2016) found that less obtrusive enhancements forms (bold printing) were disregarded. He assumed that they were less effective in drawing the learners' attention to the target forms. He also

[78] This goes not just for the data in my study, but is usually assumed to be the case, see Literature Review for details.

suggested that, “as L2 learners have years of experience with these types of enhancement”, they “may have become insensitive to them (i.e. selective attention and inattention)” (p. 85). This might be true for the Danish learners as well. The interviews offer some further insight into why bold-printing did not lead to the bold-printing group performing significantly different from the control group more often. Several interviewees mentioned that bold-printing was “confusing” and “not helpful” (e.g. 5.3.4, O’Donnell, 2012 for similar findings). This means that while highlighting the words may have caused focus on form, its effect was not strong enough to lead to distinctive results in the facets of vocabulary knowledge assessed in the VPT. Without clear instructions as to how to use them, bold-printed target words did not result in measurable vocabulary knowledge gains. Instructions may be helpful, because, as previous TIE-research has demonstrated (Robinson, 1997; Sharwood Smith, 1991; 1993), TIEs are more an implicit than an explicit attention-focusing device (Qualitative Findings). As such, their underlying purpose may not always be transparent, even to learners with some prior knowledge of the target form. Several participants in White’s (1998) study reported, via a post-treatment debriefing questionnaire, that they were not certain about the purpose of the enhanced forms. For learners with little prior knowledge of the target form, TIEs could even be more of a puzzle. Participants in the Leow (2001) study provided, among other things, the following comments in their think-aloud protocols: ‘I don’t know what that is (means)’, ‘I don’t understand these underlined ones’, ‘I am not sure’, ‘I don’t know why this is underlined’ (p. 502). Likewise, as shown in the protocols from Alanen (1995), some participants who reported noticing the use of italics had not considered a reason for its use. Several statements in my interview data reflect these findings closely.

At the same time, some of my interviewees (e.g. Stina in her first and second interview) also mentioned how useful the bold-printing had been, but for purposes other than those intended, for instance as a visual scaffold for answering the reading comprehension questions, rather than for vocabulary learning. This shows how important it is for teachers to carefully consider the purpose of enhancements. Simard (2009:133) therefore claimed that enhancements should not be based only on teacher preferences as “a careful selection of typographical cues might really improve the quality of instructional material” and could strongly influence learning outcomes regarding vocabulary acquisition, reading comprehension or other purposes (see also Balan, 1989).

To answer the first research question, I also wanted to explore whether the other two types of enhancement, glossing (C2) and the combination of glossing and bold-printing (C3), had a measurably stronger effect on vocabulary acquisition than Condition 1, and if so, why. The overall finding was that the enhancements containing lexical aids (glosses) were more effective than those containing typographic help only (bold-printing). This finding reinforces

outcomes from previous studies which reported that guessing meanings from context, as is necessary in no-gloss conditions, may not be as effective as reading while consulting glosses (e.g. Ko, 2012; Simard, 2009) and that glossing significantly facilitated learning of the target language (Bowles, 2004; Hulstijn et al., 1996; Jung, 2016; Watanabe, 1997). In my study, the superiority of glossing was obvious for instance in the results of the ordinal logistic regression calculation, which showed that only the two glossed conditions were significant predictor variables for the VPT scores (Table 22). Similarly the results of the pooled data analysis, which only recorded a significant effect on all three assessed vocabulary knowledge types for glossed reading conditions, but not for bold-printing (Table 19). The pooled data further revealed that significant group differences existed between the least (C1) and the most invasive form of TIE (C3) for all three assessed types of vocabulary knowledge. In other words, enhancements containing glosses yield superior vocabulary test scores according to different statistical measurements.

The appropriateness of these findings is supported by the interview statements regarding all three TIE-types. While Condition 1 was described by interviewees as unclear and confusing, Condition 2 and 3 were seen as transparent and helpful.

The largest correlation coefficients between vocabulary acquisition and TIE-type were found for Condition 3 (Table 20). The statistical analysis showed that Condition 3, the combination of bold-printing and glossing, led most often to significant effects on VPT scores, compared to the other enhancement types and the control group. This is in line with the outcomes of several vocabulary related TIE studies (Jourdenais, 1998; 2001; Kost et al., 1999; Min, 2008; Pacheco, 2004; Peters, 2012; Simard, 2009; Yoshii, 2006; Zandieh & Jafarigohar, 2012) and also grammar-focussed TIE studies, for instance Izumi (2002) and Williams (1999). I already discussed various reasons for these findings (Quantitative Findings). All of these studies confirmed that a combination of different enhancement forms led to results superior to those of simple enhancement types.

Linking the VPT scores with the interview data explains why this enhancement type was so advantageous. The interview data seem to both confirm and challenge the above-mentioned findings. On the one hand, Jeppe confirmed that the ‘feedback’ and the related ‘checking effect’ inherent in glosses paired with bold-printed target words influenced the way in which he worked with Condition 3. He explained that he deliberately interrupted the reading flow, stopped in the middle of a sentence and checked the definition in the margin and then returned to the text (1.45). At the same time, he felt worried that this might not be the optimal way of working, as it interrupted the reading. Other interviewees reported that they had felt compelled by the enhancements to go to the margin to check the definitions (e.g. Jeppe 1.45, Fie 1.34, Stina 3.6). This confirms that combined forms of enhancement encourage learners

to focus on the target words in ways that are likely to initiate deep processing (decontextualizing the word form from context, checking the word's meaning in the margin, processing and then re-contextualizing the given information) – probably more than the other forms of TIE investigated in my study. In contrast to this stands Fie's statement, in which she explained the shortcomings of the glossed conditions. She pointed out that glossing, although convenient, deprived learners of the chance to process the words and come up with their own conclusions (Fie 1.36). Despite the fact that only a one student referred to this, it seems worth mentioning here for two reasons. Firstly, her statement highlights an important dilemma regarding vocabulary acquisition, namely that glossing provides immediate feedback and thus truncates the hypothesis formation in the vocabulary acquisition process (Cho, 2010). Secondly, her concern is interesting, because it is in clear contrast to the results reported for the glossed reading conditions. Even though, or rather because, glossing interrupts the reading flow, provides the 'answers' and learners seemingly do not have to do the deep processing themselves, the results of the glossed conditions are still superior to those recorded for bold-printing. As proposed by previous research (Rott, 2007; Simard, 2009), the obtrusiveness of glossed TIE-types is beneficial rather than hindering, as it works *for* deeper processing through the feedback contained in the glosses, so that the involved switching cost is not too high in terms of its effect on vocabulary acquisition. Similar findings were reported by Cheng and Good (2009), who found that, even though glosses in the margin were distracting, they still contributed positively to the acquisition process. They assumed that this was because the glosses made the learners focus on processing word meaning. Potentially, however, deciphering the vocabulary may have taken up so much of the learners' cognitive resources that there were only limited mental resources left to dedicate to reading comprehension. Therefore the issue of 'switching cost' must be considered in the context of reading comprehension. Switching cost is a concept that emerged from reading research. It explored the question of whether juggling limited cognitive resources, which is likely the case when readers move from text to margin, leads to greater inefficiency and lower productivity when completing a comprehension task (Rubinstein et al., 2001). This is relevant in the context of concurrent processing vocabulary acquisition and reading for comprehension, as, while the obtrusiveness of the glosses seems to serve vocabulary acquisition positively, it might hinder comprehension. As my treatment did not contain a measure of reading comprehension, my data does not shed light on whether vocabulary acquisition occurred on the expense of comprehension. Previous research has delivered inconclusive results (Cho, 2010; Han et al., 2008; Jung, 2016, Winke, 2013).

Another issue explored in this first research question was the link between TIEs and incidental and intentional learning. The combination of quantitative and qualitative methods helped to

illuminate this matter. The statistical analysis allows for comparisons between the scores recorded in the first (assumedly incidental vocabulary learning) and subsequent sessions (assumedly increasingly intentional vocabulary learning), and the learner behaviour described in the interviews reveals whether any learning that did take place seemed to be incidental or intentional. As explained in the Methodology Chapter, in my study the factor 'session' is understood as an indicator of whether vocabulary learning was approached incidentally or intentionally.

The statistics show that outcomes regarding the effects of TIEs on vocabulary acquisition are different from session to session (Tables 11, 14, and 17). Even though some of the significant group differences appeared between the same groups in Session 1 and Session 2 (C1-C3 for passive recall; C0-C3 and C1-C3 for passive recognition), and between Session 2 and Session 3 (C0-C3 for passive recall), several findings differ from session to session. These findings reveal no clear pattern that would allow conclusions regarding the incidental/intentional learning distinction. In the second and third reading/testing session, the control group scored significantly different from Condition 3 for passive recall (Tables 14 and 17). This shows that even though participants were assumed to paying increasing attention to learning the target words (intentionally) the TIEs upheld their effect. In other words, the combination of glossing and bold-printing had an impact on passive word meaning recall that was significantly stronger than that resulting from encountering the target words without enhancements.

This occurred even though the participants were increasingly aware of procedure and purpose of the reading/testing process. The interviews showed that, in contrast to the first session, an intentional approach to word learning was taken in the second and especially in the third session. Here, task- and test familiarity led to a more economical working style and a heightened awareness of the target words. Stina, for example, started reading the text in Session 3 only after having read all target word definitions in the margin. In addition, she also stopped again while reading a second time, "just to make sure" that she knew the words (3.18). Statements by Silas (3.3), Fie (3.10), and Jeppe (3.7) describe a similar behaviour. As this shift from incidental to intentional learning is so clearly revealed by the interviewees, it is likely that at least some of the other (un-interviewed) participants used a similar learning approach and that the results of the statistical analysis reflect its impact. The logistic regression calculation suggests that this is the case. It identified the coefficients of the assumed predictor variables (Table 22) and showed that in the last round of reading/testing, the factor 'session' significantly predicted the outcomes of the VPT for active recognition and passive recall. Moreover, the students who had worked with Condition 3 significantly outperformed those in the control group in the passive recall task in Session 2 and 3. This reveals that even when the target words were approached intentionally, the meaning recall task was so demanding

that the help the TIE provided was needed to outperform those participants who had worked with unenhanced texts. Accordingly, it can be concluded that when the purpose of providing enhancements is to boost the ability to recall meaning, then meaning focussed enhancements like glosses are useful to guide learners' attention to these words.

Because of the increasing test awareness and the resulting attention participants paid to the target words, the line between incidental and intentional learning behaviour in my study is a relatively clear one. For this reason the results of the three sessions were analysed separately as well in a pooled data analysis. Regarding the relationship between the two learning approaches and use of TIE, however, the line seems blurred. It is not quite clear how either incidental or intentional word learning was linked to TIE use as the statistical results did not show a clear pattern (see above). This is linked to findings from research on incidental and intentional learning, which established that it might be theoretically difficult to maintain the incidental/intentional distinction as the two concepts are closely related, while recognising that it might be essential to maintain the distinction methodologically (Barcroft, 2004; Gass, 1999; Hulstijn, 2001; Huckin & Coady, 1999). The data analysis in my study confirmed that incidental and intentional learning can be difficult to distinguish and suggests that a number of factors may be influential when trying to distinguish them. This is in line with Huckin and Coady (1999:190), who pointed out that, for example, incidental learning was "never entirely incidental," as the learner must pay at least some attention to individual words. However, again according to Huckin and Cody (1999) and as confirmed in my study, the amount of attention and the amount of learning varies according to a number of factors, including context, type of attention, and task demands.

Finally, the perplexing relationship between incidental and intentional learning is an unresolved TIE-related problem with regards to how the finding can be understood that Condition 3, the most obtrusive form of TIE[79], yielded the highest scores for both receptive and productive vocabulary knowledge. The reading task was geared towards an incidental target-word approach, i.e. word learning was supposed to be treated as a side aspect while the main focus was supposed to be on reading comprehension. It seems obvious that an amplified focus on form and meaning (like in Condition 3) would lead to higher test scores. To return to the question raised in the Literature Review of whether more attention would lead to more learning (Schmidt, 2001), it must be concluded that this was the case. As raising salience is likely to raise the attention learners pay to words, they will become more explicit and more likely to be learned intentionally rather than incidentally. However, this seems to be in contrast to the original idea of enhancements as an implicit, subtle learning device aiding incidental

[79] A combination of bold-printing and glossing.

vocabulary learning. Traditionally, TIEs are used to further word learning as a by-product of reading, not the focus of attention. However, with an intrusive enhancement form such as Condition 3, I would argue that lexical processing should be considered similar to explicit vocabulary learning. Both my quantitative and qualitative data suggest that working with such combined enhancements means that they are used different from the original concept of textual enhancements, as an implicit vocabulary learning or reading help. This suggests that in TIE-related research the terminology regarding incidental and intentional learning needs to be more carefully used. If vocabulary studies use combined forms of enhancement, the research design should try to assess whether the occurring acquisition processes are to be located more on the incidental or intentional side of the learning spectrum. The finding that using combined enhancements may lead to more explicit vocabulary learning is relevant also as it relates to the Involvement Load Hypothesis (Literature Review), which was initially linked to my study (Introduction). In the original involvement load study (Laufer & Hulstijn, 2001) combined enhancements (bold-printing + glosses) were used as part of a task design illustrating a specific level of involvement load. However, the cognitive processes linked to using these enhancements while reading were regarded as non-existent. They were not considered in the calculations determining the level of cognitive involvement load. According to my findings, this is problematic. Using enhancements, and in particular when target words are enhanced twofold, does carry a cognitive involvement load which influences the level of cognitive involvement load triggered by a task. I would therefore suggest that this should be accounted for with a weak *evaluation* factor (involvement index 1, Laufer & Hulstijn, 2001). Laufer and Hulstijn clearly formulated their hypothesis as just that - a hypothesis. Even though they provided empirical evidence for its claims (Hulstijn & Laufer, 2001; Keating, 2008; Kim, 2008), they regarded involvement load as a construct and the hypothesis as explorative. In the subsequent research literature, however, it is often referred to as if it was a fact/an established theory (e.g. Folse, 2006; Gass and Selinker, 2008; Joe, 2010). This is despite the fact that my research shows that the underlying assumptions of this hypothesis need more consideration.

6.3. Research Question 2: “Does Textual Input Enhancement Have A Differential Effect On Different Types Of Word Knowledge?”

With regards to this research question, it is difficult to draw clear conclusions from the data analysis. In particular the results concerning part C of the VPT (‘passive meaning recall’) are not fully clear, as there is no recall-counterpart the scores could be compared to (Nation,

2001)[80]. However, the results of my study suggest that retrieving a correct L2 word meaning when exposed to the L2 word form displays active/productive vocabulary knowledge rather than, as Laufer et al. (2004) suggest, passive/receptive knowledge. In the light of my data analysis with regards to the different tested vocabulary knowledge types, I would claim that asking learners to retrieve L2 word meaning from a given word form (recall - part C) can not necessarily be regarded as 'passive', just as it is not necessarily clear why selecting the correct L2 word form to fit the provided L2 word meaning should be regarded as 'active'. The fact that part C of the test was often perceived as more difficult than the other two parts and very often led to significantly lower VPT scores, suggests that this part of the test measured active/productive knowledge, which reflects a higher level of the vocabulary acquisition continuum than passive/receptive knowledge (see Literature Review).

Further relevant with regards to research question two, Laufer et al. (2004:218) found that the results of the two recognition tests were "indistinquishable from one another in terms of difficulty". The outcomes of my study confirm this. In my study test parts B and D tested word (form and meaning) recognition. My findings show that the order of the test parts as suggested by Laufer et al. (2004) and as based on an assumed difficulty hierachy was not necessarily plausible. The scores from part D (passive meaning recognition) were not always lower than those from part B (see e.g. Tables 9, 12, 15). However, as the order of the test parts in my study was different from Laufer et al.'s, other factors, such as test fatigue, may play a role. In addition to this, the proficiency level of my tested students is somewhat lower than the participants of Laufer et al.'s (2004) study.

In research question two, I also wanted to investigate whether, when learners focus simultaneously on both form and meaning while reading, or just one or the other at a time, TIEs are a way of aiding or guiding this type of focus. The interviews do not offer systematic insight into this question, but on the other hand clearly showed that meaning recall was more challenging than meaning recognition (e.g. Stina 1.20; Stina 3.39-43, Jeppe 2.27, Jeppe 1.63, Jeppe 3.39). The interviews indicate that the glosses most significantly influenced in the recall tasks. Jeppe makes it clear that he used glosses to establish a form-meaning link (Jeppe 1.99). However, even bold-printing led to noticing of word form and meaning.[81]

The statistical analysis revealed diverse results, which, however, could not be interpreted clearly. This may be due to the fact that the levels of test-awareness varied from session to

[80] The 'active meaning recall' was eliminated from the test setup after the pilot study.
[81] Eliminating the test part which measured active form recall (after the pilot study, see Methodology Chapter) because it was too difficult for my study's participants, tacitly presuppsoses that this would have been the most challenging part.

session, but it is probably also linked to the fact that form knowledge was only tested in one ('active') recall task, while meaning knowledge was measured in in two dimensions, 'active' and 'passive'. For example in Session 3 (see Table 17), the TIEs did not have a significant impact on the parts of the test that measured form and meaning recognition, only on meaning recall. These outcomes are similar to those by Rott (2007a), Watanabe (1997) and others. How these results link to the participants' vocabulary learning focus (on form, on meaning, or on both) is uncertain. It seems that Nation's (2001) warning, not to mix recall and recognition items in tests was warranted. It is not possible to get a balanced view of how the TIEs influenced either form or meaning knowledge, as the active/productive vs. passive/receptive dimension would have to be taken into account as well. These two dimensions interact, i.e. the form-meaning link of a word may be known productively, but how word behaves grammatically may only be known receptively. It would have been helpful to be able to differentiate between word form and meaning as aspects of word knowledge on the one hand, and the receptive/productive dimension as different forms of mastery on the other (Schmitt, 2010). Unfortunately, the interviews did not provide systematic information about either of the two concepts, so that the results of the statistical analysis cannot be assessed in the light of such additional qualitative data.

It would have been interesting to learn more about how well the different target words were acquired (depth of knowledge) or also the probable connection between recognition/production and the concept of incidental/intentional learning. Peters et al. (2009), for example, had found that test announcement (intentional learning) helped word *recognition*, but did not result in higher *recall* rates. However, here again, my interview data does not allow reliable conclusions.

6.4. Research Question 3: "Which Task-Related And Learner-Internal Factors Emerge As Relevant For Vocabulary Learning From Reading Enhanced Texts? In What Way Do These Have An Impact On How Input Enhancements Are Used In Reading Tasks?"

Many reasons might have initiated the vocabulary learning assessed in the VPT and investigating the effect of the different types of TIE covers only part of the picture. Six of the potentially influential aspects were included as variables in the measurements (session, gender, school, language background, and text). The statistical analysis (section 4.6.3)

showed that several of them did significantly affect vocabulary acquisition. This research question was intended to examine the impact of these factors and explore which other elements were relevant for the outcomes. The interviews revealed that TIE-related instructions, test (type), text difficulty and proficiency were particularly influential. These aspects were discussed in detail in chapters 4 and 5.

It is clear from the interviews that the task instructions were an influential factor regarding task-completion and language learning. They helped to explain some of the TIE-type related outcomes of the statistical analysis. For example, they show why bold-printing led to scores much lower than the glossed Conditions (Condition 2 and 3). In my research, the instructions deliberately pointed out only that the enhancements were there to help with the reading. However, according to my experience as a teacher, most textbooks and also most educators do not usually give TIE specific instructions or only sparsely. Statements in my interviews as well as findings from previous research showed that all types of enhancements can cause problems if they are not linked to instructions. One of my interviewees, when reading a text with bold-printed target words without instructions, could not see the purpose of the bold-print and therefore decided to ignore the enhancements (Silas 1.58). Runa "could not understand why the words were 'black' [bold printed]." She "did not know what to do with them" (2.22). These findings are similar to those of for instance Alanen (1995), Leow (2001), and White (1998), whose subjects had, in reaction to various types of enhancements, reported puzzlement, uncertainty of how to work with the enhancements, and said that they had consequently ignored the enhanced words. O'Donnell (2012:557) was surprised by the "various forms of misapplication of marginal gloss information" that she found in her data and assumed that these were, in addition to matters of gloss formatting, related to the learners' erroneous interaction with the glosses. Providing clearer instructions might have been a solution. Similarly, the fact that in my study Runa made the above-mentioned statement after having read a text enhanced in Condition 3 (bold-printing and glossing) suggests that clear and salient instructions might be beneficial for all kinds of enhancements, even for types such as glosses that contain a seemingly obvious purpose and clear link between the target item and its explanation.

However, Sharwood Smith (1991, 1993) pointed out that TIEs were more an implicit than an explicit attention-focusing device and that, accordingly, their purpose may not always be transparent. While being ignorant of the purpose of a teaching/learning intervention might not generally be a problem, my data confirmed the findings of previous research that, when it comes to TIEs, not providing instructions can be problematic and can get in the way of

learning. If the *need*[82] of the TIEs is not obvious to the students, this may affect task completion, motivation and thus also language processing. This is because enhancements, according to my findings, usually still are sufficiently explicit to be confusing for learners. O´Donnell (2012:546) argued that questions regarding gloss type, albeit important, were "of little consequence if readers are not using the glosses as intended". Therefore, if TIEs are used with a learning purpose in mind, providing instructions is recommended. This presupposes that the provider of the enhancements is aware of the purpose.

Peters et al. (2009:143) found that "students allocate their attentional resources in function of the specificity of the task they have to perform". This means that clearer instructions linked to target words may lead to processing. Instructions could help to point out potential problems (e.g. general confusion, ignoring the unknown words). My findings suggest that even in their ordinary foreign learning classes, students might ignore target word enhancements if no instructions are provided.

It is possible that providing instructions would shift learning from incidental to a more intentional approach, but not necessarily so. The learning approach still depends on the learner's interaction with the enhancements.

The second issue related to research question three it that of text difficulty. In both the qualitative and qualitative data analysis, it became obvious that text-inherent factors contributed to the outcomes. The chosen texts were different from each other not only regarding their topic, but also with respect to difficulty. This influenced the VPT results and also student behaviour. Whether or not words are perceived as '*need*ed' for task completion and accordingly how TIEs are used is linked to text difficulty. This was confirmed in O'Donnell (2012), where participants used the input enhancements much more extensively in the text that was categorized as being rather more difficult. A similar tendency emerged in my interview data, when the interviewees confirmed that the TIEs had been very helpful when reading the comparably more difficult Texts 2 and 3 but less important for reading the easier Text 1 (e.g. Jeppe 1.37; Silas 2.2.6; Fie 1.34). When 'text' was included as a potentially influential factor in the ordinal logistic regression, it emerged that it did predict the outcomes of the VPT (in all three parts of the VPT and with coefficient values between -.53 and -1.2), i.e. Text 2 and 3 did contribute strongly to the test outcomes and led to significant lower scores than Text 1. The interviews clearly support these results. Text 1 was regarded as easier than the other two texts (e.g. Stina 2.4, Jeppe 2.2). O'Donnell's research (2012) also links the use of TIE to text difficulty. She found that the degree of benefit that readers derive from marginal glosses is

[82] If they do not find the word essential for completing the task, see section 2.3.5 about the Involvement Load Hypothesis.

directly related to the level of text difficulty, i.e. gloss usage increased with rising text difficulty. Problems arose because the learners either did not know the meaning of or did not understand how the words or phrases related to the storyline. O'Donnell also found that the efficacy of using TIEs is directly linked to general reading performance. In her study, those readers with better comprehension referenced the marginal glosses twice as often and used the information provided in the margins almost twice as successfully as readers with poorer comprehension. Cheng and Good (2009:128) suggested that if text difficulty was beyond learners' comprehension capacity, glosses would not "efficiently facilitate subjects' understanding of the whole passage or allow them to acquire the target words". Jacobs et al. (1994) also mention too low/high text difficulty as one reason for why their TIE treatments sometimes did not lead to effects. According to my interview data, it appeared that text difficulty should be considered as one factor influencing the measured effects of the TIE treatments.

Finally, by looking at both types of data I understood better how the testing situation had had an impact on the outcomes of the VPT. In the interviews, test format emerged as a test-related problem. Regarding the two multiple-choice parts of the VPT (B and D), it became obvious that the type of assessment influenced the way the students handled the test. Fie and Jeppe reported that they had found part B and D easier than C and all five interviewees either directly said or implied that, rather than completing these test parts with a focus on vocabulary, they had used logical thinking and exclusion strategies (section 5.6.4). These statements support previous research findings, which revealed potential test-related validity problems, especially with multiple-choice formats (e.g. a 'negative suggestion effect' through exposure to wrong answers, guessing correct answers irrespective of knowledge, see for instance Roediger & Marsh, 2005; Stewart, 2014; Stewart & White, 2011). Therefore, it is likely that test format contributed to the high scores recorded for both active and passive recognition in all three sessions (see Tables 9, 12, 15). The interviewees' statements suggest that the multiple-choice test format might be one reason why no significant distinctions were found for recognition between the TIE groups in the third session (Table 17). Beyond the fact that recognition is probably easier than production, it is likely that the participants' test strategies overrode the effect of the TIEs in these instances. Yoshii (2006) found similar test format effects in a study which investigated the effect of different types of L1 and L2 glosses (text-plus-picture versus text-only) on vocabulary learning. He had used a definition-supply test, which included pictures and a multiple-choice recognition test without pictures. The results differed according to test type. Yoshii concluded that test format, i.e. whether or not "learners were able to see the multiple choices and use them as hints for recalling the meanings" had a stronger impact on vocabulary learning than the pictures and thus overrode their effect (p. 95). In my study, different types of tests were applied also to counteract the impact of such test effects on the

outcomes, but it has to be acknowledged that these still influenced learner behaviour and perception.

6.5. Chapter Summary

The main objective of this chapter was to discuss where the interview data provided a better understanding of the statistical findings in relation to my three research questions. With this, I follow recent studies in the field in which different methods were applied to illuminate TIE-related issues (e.g. Bell & LeBlanc, 2000; O'Donnell, 2012; Winke, 2013). In the following I show that in my study the different data sources sometimes contradict and sometimes confirm each other.

Triangulating methods was useful for understanding the various effects of the different TIE types. The interviews disclosed different aspects, which explain why bold-printing led to much lower scores than glossing (C2) or Condition 3. With this, my findings give insights that exceed what can be assumed by common sense, i.e. that providing word meaning (as in glosses) aids vocabulary intake more clearly than highlighting the form of the word only.

Contrasting the findings from the quantitative and qualitative data analysis proved also advantageous for understanding how TIE-use links to incidental and intentional approaches to vocabulary learning. The results elicited from analysing the VPT scores suggested that the approach continuously shifted from incidental in the first testing session to a more and more intentional approach in the second and last session. Whilst this influenced the test outcomes, test familiarity and other factors also played a role. The interviews showed that text difficulty, testing and the nature of the TIE-related instructions did have an impact.

Unfortunately, the qualitative data did not contribute much to a better understanding of the second research question, as they contained little useful information regarding the impact of TIEs on form and meaning and recognition and production in vocabulary learning.

In the following chapter, I conclude this thesis by discussing the wider implications of my study, their significance of their pedagogical application and their importance for future research.

7. CONCLUSIONS

7.1. Project Summary And Major Findings

This chapter concludes my study by revisiting the major findings, highlighting implications for language teaching and discussing the limitations of this project. Lastly, I present issues important for future research.

Summary

In this study, I set out to investigate phenomena that I had encountered in my teaching experience. I wanted to investigate whether using word enhancements with the purpose of aiding vocabulary learning had a measureable effect on how well these words were acquired and what types of learning processes determined this outcome. I was further interested in finding out which other task- and learner internal factors played a role.

Major Findings

The collected data allow insights into the effects of TIEs on vocabulary learning and the ways TIEs are used by language learners. I summarize the most important findings under the three research questions.

Research Question 1: What are the immediate and long-term effects of different types of textual input enhancement on incidental and intentional learning of EFL-vocabulary from reading tasks?

I found that using textual input enhancements (TIEs) to highlight previously unknown vocabulary in reading tasks did aid form/meaning recognition and recall of these words. A strong correlation was found between the use of TIEs and vocabulary acquisition. The effectiveness of textual enhancements for vocabulary acquisition depends on their type and the form of vocabulary knowledge assessed (e.g. passive recall, active recognition). Lexical enhancements containing glosses were more effective in aiding word learning than typographic enhancements that only highlight word form. This was confirmed by the different analytical approaches to the data, for instance the data collected in different sessions, and also when I analysed the data as a whole. Enhancement forms containing glosses were found to aid vocabulary acquisition to a degree that even the ‘instant feedback problem’, the fact that

they immediately fill vocabulary knowledge gaps and thus render further processing unnecessary, is overcome. Enhancing target words with a combination of glosses and bold-printing (Condition 3) was the strongest of the three investigated enhancement forms in its effect on vocabulary acquisition. My data showed superior results for this reading condition compared to the control group, the non-glossed enhancements and often also to the gloss-only condition. These findings shed light on the concepts of 'switching cost' and 'obtrusiveness'[83]: They suggest that the obtrusiveness of this enhancement form is beneficial for processing new words rather than being an impediment to learning. This is due to the attention paid to the target words and the fact that it encourages deep processing behaviour. In contrast to that, I found that the effect of highlighting unknown vocabulary only typographically is almost negligible for aiding vocabulary learning.

The finding that the combined form of enhancement influenced vocabulary acquisition most clearly links to the complex relationship between incidental and intentional learning when it comes to TIE use. As it must be assumed that working with such intrusive enhancement types actually promotes intentional rather than incidental learning, my data casts doubt on previous studies which worked with similar types of enhancements. In particular it also suggests that the theoretical concepts constituting the measurements of cognitive involvement load proposed in the involvement load hypothesis (Laufer & Hulstijn, 2001) need to be reconsidered.

The data analysis revealed that the repeated reading/testing procedure initiated a shift from an incidental to an intentional learning approach between the first and the subsequent sessions. This is reflected in a heightened awareness of unknown vocabulary, and resulted in the use of different test-taking strategies. Accordingly, it became obvious that repeated testing has a noticable impact on test outcomes, i.e. that research design partly determines the findings. This had to be taken into account in my study, some of my outcomes had to be treated cautiously. At the same time, this points to problems with similar previous and future research, which might suffer from the same weakness (e.g. Eckerth & Tavakoli, 2012; Folse, 2006[84]).

Research Question 2: Does textual input enhancement have a differential effect on different types of word knowledge?

A single exposure to enhanced L2 vocabulary was found to result in limited but selectively

[83] 'Switching cost', i.e. moving between text and margin may strain a learner's cognitive resources, and 'obtrusiveness', i.e. to which extent glosses interrupt the reading flow; for details see Literature Review.
[84] Even though the treatment procedures in these studies varied slightly from session to session, they were still so similar that it is likely that the described effects influenced participant behaviour and thus the outcomes.

significant acquisition of word knowledge, predominantly in form/meaning recognition. This confirms findings from previous research into incidental L2 vocabulary acquisition from reading and suggests that multiple exposures to the same words in different contexts would be necessary to consolidate more advanced knowledge of those words (e.g. meaning recall). The interviews confirmed that reading a glossed word is likely to leave a memory trace, i.e. that externally imposed salience and incidental processing do lead to the acquisition of some forms of initial word knowledge.

Two important findings related to TIE-type and its impact on specific word-knowledge-types. One is that the type of TIE with overall superior effects on word learning (C3 = bold-printing + glossing), was most beneficial for meaning recall[85]. This suggests a link between enhancing word form together with providing meaning help and students' ability to produce this type of word knowledge. The second finding is that the strongest impact of TIEs on the vocabulary post-test (VPT) results was measured for the recall part of the vocabulary test. This indicates that while TIEs may be useful for developing the abiltity to recognize word form or meaning when reading (assessed in part B and D of the VPT), they are vital for enabling learners to acquire word knowledge at the level of recall. However, my data analysis suggests that the categorisation of the different test parts, in particular the 'active' vs. 'passive' distinction, as suggested by Laufer et al.'s (2004) CATSS design was ambigious.

As mentioned above, the effect of bold-printing alone was comparatively weak. However, significant effects of bold-printing were measured for 'passive' meaning recall (part C of the VPT).

Research Question 3: Which task-related and learner-internal factors emerge as relevant for vocabulary learning from reading enhanced texts and how do these affect how learners use input enhancements in reading tasks?

Unlike most previous TIE studies, I integrated qualitative data into my study and was able to shed some light on ways in which learners interact with TIEs and how their behaviour influences vocabulary learning. These findings not only enlightened my quantitative data analysis, but also shed light on the outcomes of previous research.

The learners noticed and perceived the different types of TIEs differently. The TIEs were regarded as more or less useful and were used in many ways beyond the purpose of vocabulary acquisition and reading comprehension (e.g. as a visual scaffold in text

[85] Learners were presented with the target word form and asked to paraphrase their meaning or to provide a synonym.

navigation). The interviews showed that the use of TIEs is highly individual, and that language learners do not necessarily use TIEs for language learning, as they are primarily interested in quick task completion, and less in language processing. I also found that, beyond the TIEs, several of the recorded other independent variables, such as 'session' (whether the data was collected in the first, second or third testing/reading session) and 'text' (which of the three texts was read), were factors that had a significant effect on the outcomes of the VPT.

Another finding relates to the impact of participant awareness in my study ('Limitations'). As a consequence of repeated testing, test participants became increasingly aware of the test procedure and purpose and this has affected and potentially limited the research findings. This is an important concern, because it is likely to have affected previous language learning studies also. Future research working with repeated measures designs need to factor this in, and the outcomes of previous studies may have to be reconsidered in the light of this concern.

My data analysis further confirmed that the motivational *need*-factor of the involvement load hypothesis (a learner's *need* to pay attention to the L2 input) is reflected in learner behaviour when dealing with TIEs. If the target words are relevant and meaningful (i.e. needed) for task completion and perceived as such, learners are likely to notice and process them. Otherwise, they are more likely to be ignored. This perception, however, I found to vary individually.

7.2. Implications For Teaching Practice

Unlike most previous TIE-studies, rather than in university context, I investigated vocabulary learning in schools, the context in which most English language teaching occurs (secondary school). Moreover, my participants are 'intermediate' learners, arguably quite advanced learners. As a result, my findings are most applicable to foreign language teaching practices for intermediate learners in secondary school contexts, but relevant even to university courses.

My study identifies a number of issues concerning TIEs that teachers should be aware of. These issues concern how the enhancements should be presented, whether and which instructions should be given, and how learner behaviour can affect learning efficiency. All of these findings suggest that using enhanced reading materials productively calls for careful planning.

How TIEs should be used and presented to learners depends on the teaching purpose and on the learner's preferences. Concerning the latter, my interviews showed, for instance, that

some learners habitually produce textual enhancements to support their reading and vocabulary learning. My interviewees' statements made it obvious that such learner-created salience builds on learners engaging interactively with texts and unknown words, as they meticulously customize such enhancements to their own learning needs. Therefore, this technique should be usefully encouraged in all learners.

The second conclusion relates to the former of the above mentioned points, the teaching purpose. As TIEs are more an implicit than an explicit attention-focusing device, they are seldom linked to specific instructions. I found, however, that instructions play an important role with regards to how successfully TIE are used for either language learning or reading comprehension. Thus, my findings confirm the findings of previous research (e.g. Winke, 2013; Zandieh & Jafarigohar, 2012). However, to determine which types of instructions exactly are appropriate, teachers must think about why they want to use TIEs, i.e. whether the enhancements are primarily aimed at fostering reading comprehension or vocabulary learning, or a combination of these two. I found that generally, appropriate instructions for the use of TIEs are recommended, but especially necessary when the focus is on vocabulary acquisition. This is in line with Rott's (2007a) assumption that readers without instructions to do otherwise, do not direct their attention to lexical form but focus on comprehension as a default. Without instructions, learners are uncertain about the use of the enhancements, which is counteractive to focussed learning and the teachers' intentions in providing TIEs. Instructions help learners to distribute their attention purposefully.

Accordingly, if reading comprehension is the primary goal of using TIEs, my data shows that TIE-related instructions are not essential. However, even though providing instructions can never guarantee that learners follow the intended path, they can maximize the effect of TIEs for attentive students or learners in need of guidance. Choi (2016:16) recommended that when the primary purpose is on helping learners make sense of the text, teachers "should take great care in implementing textual enhancement". He suggests encouraging learners "to read the same text twice: first for content and overall comprehension and a second time for useful language".

My data further suggests that enhancement forms containing glosses are preferable over other types of enhancements, especially if the purpose is vocabulary knowledge acquisition beyond the level of recognition. If the purpose of using TIEs is to foster the ability to recall form or meaning of a word, then the effect of enhancements of the type chosen in my study is not strong enough and should be combined with other means of raising salience (frequency, additional vocabulary-linked pre- and post-reading exercises). My interview data also showed that glosses need to be written with great care in order not to overload learners. Gloss-

definitions must be appropriate for the learners' proficiency level. If they are not able to understand either their purpose (see above) or the wording, they are counterproductive.

Another one of my findings is important for understanding how TIE use relates to vocabulary learning. I found that, besides factors such as, for instance, TIE form or obtrusiveness, learner-internal variables had a very powerful effect on the learning outcomes. This confirms Sharwood Smith's (1993) claim that internal enhancements may dominate over external enhancements as they may or may not overlap. If the purpose of TIEs is enhancing learners' knowledge of the highlighted vocabulary, words that are clearly meaningful and relevant for task completion should be targeted. This is necessary to cater for two types of learners. Some learners primarily tend to words that are perceived as 'needed'. My interview data and previous research suggest that this 'need' factor is an important indicator of further word processing (e.g. Hulstijn & Laufer, 2001). The other type of learner approach texts enhanced with glosses habitually by always reading all glosses. Guiding those learners specifically to lexical items that are relevant helps them to work more effectively. However, anticipating such learner perceptions requires a high awareness of, for instance, task structure, teaching purpose, learner preferences, and learner strategies on the side of the teachers.

7.3. Limitations

In hindsight, it is clear that some shortcomings of this study were based on the methodological choices I made when approaching the research questions. This section describes some shortcomings in my methodological approach and limitations regarding the findings of this study.

I was able to answer the three research questions to a large degree. However, some aspects remained unclear.

For instance, I was not able to uncover any long-term effects of the TIE use assessed in my study. Too few students were available to be retested, so that the data sample of the delayed post-test was too small to obtain statistically reliable information. Using alternative types of vocabulary knowledge tests (e.g. asking participants to integrate target words into a writing task, word recognition tasks, and interactive reading tasks) could have been alternatives to retesting. These might have shown whether durable learning could be traced back to being exposed to the target words under different reading conditions. Further, to solve the statistical problem of zero inflation (section 4.7), it might have been possible to split up the data again and compare the TIE types one by one, instead of in one model.

The best alternative, however, might have been retesting, i.e. to give out the vocabulary post-tests again, in an unobtrusive way, e.g. ten words at a time or, alternatively, where students would have read one text with five unenhanced, five glossed, five bold printed words, another student another text with different types of enhancements, etc. whenever there was time. Thus, it would have been possible to compare the test scores internally (for each student), whithout having to rely on full classes attending a proper testing session. Proceeding in this way, may have made it possible to gather delayed-post-test data of far more students independent of class attendance during the testing sessions. However, unfortunately, this would have required planning this set-up and integrating it into the design of the study from the very start.

My data also provided only little clear information on the differential impact of the different types of TIE on the two assessed types of receptive knowledge (active and passive form and meaning recognition). Even though significant effects of TIE on the two types of receptive knowledge were found, no clear-cut conclusions or generalisations were possible as the results varied strongly from session to session. This indicates that variables such as 'session', which took effect for instance in the form of test familiarity, were very powerful. Similarly, it turned out to be unfortunate to proceed with only one measure of (meaning) recall. The measure of active form recall had been eliminated after the pilot study results showed that this test part produced extremely low scores. This made it difficult to compare the effect of the TIEs on the different types of vocabulary knowledge with each other. Substituting the eliminated active recall task with another, easier version, for instance a translation task, might have circumvented this problem. However, more or longer testing sessions would almost certainly have affected motivation negatively.

My initial assumption, based on findings from related research, had been that the three different types of TIE in my study represent different degrees of effect on vocabulary acquisition from reading (C1 being least, and C3 being most effective). I believed that they initiated processing to different degrees and that weaker or stronger forms of TIE with more or less clear connections between target words and their definitions would cause processing of differing depths and accordingly lead to a word learning hierarchy. Similar assumptions were expressed by Watanabe (1997). My analysis shows, however, that many factors influence vocabulary learning from enhanced texts. While my findings suggest that there is a hierarchy (less obtrusive forms of enhancement influenced the assessed facets of vocabulary learning less significantly than more obtrusive forms), it is still not clear how far enhanced *input* is responsible for the measured *output*. A simpler research-design, investigating fewer variables would have made this input-output link more obvious.

With this simplification of the research design in mind, it had been decided that reading comprehension would not be included as a measured variable. In hindsight, this seems unfortunate. It would have been useful to see whether there was a link between vocabulary acquisition and reading comprehension, and in particular whether some of the word learning occurred on expense of reading comprehension. As has been discussed earlier (see section 2.3.1 and 2.3.2), research has shown that efforts to understand meaning in a text may take attentional focus away from processing vocabulary. In meaning-focussed tasks such as the ones completed by the participants of this study, it is likely that incidental as well as intentional processing of vocabulary withdrew attentional resources from the reading effort. Comparing the outcomes of the reading comprehension measures with the vocabulary learning test taking scores across sessions, would have shown whether any of these factors benefited from or was adversely affected by the use of the TIEs, and to which degree. Therefore, it is important to take into account that vocabulary learning was not the only focus in the investigated learning situation.

Further, the data analysis also revealed four interlinked problems regarding the chosen test-format: student awareness, training effect, test fatigue, and the type of test. These factors turned out to be more influential than I anticipated from the pilot study, and they became a concern in the analysis of the main study as they obviously influenced the findings.

The first of these problems is participant awareness: Even though I clearly explained to my participants that the tests were not assessing their competence and that the teacher would not be informed about their performance, some students still wanted to do well and became quickly alert to the objective to learn new vocabulary, trying hard to perform well and learn the new words. This was confirmed in the interviews. Therefore, the findings from sessions 2 and 3 may not reflect an incidental reading/learning scenario and the results may not reflect a natural classroom learning situation. I therefore treated Session 1 as incidental learning, but regarded the results from the following two sessions as based on intentional learning. In my interpretation of the data I acknowledged this as an influencing factor.

Similarly, a training effect and test fatigue brought about by the repeated testing was observed in the second and third testing session. Due to the repetitive structure of the VPT, some participants quickly familiarized themselves with the procedure and developed a testing ‘routine’ in which motivation decreased. They resorted to guessing and answered randomly rather than thoughtfully.

Including multiple-choice testing in the VPT was part of this problem. As has been discussed before (section 3.7.2.2), and as the interviews suggested (section 5.6.4), guessing, logical thinking, and exclusion strategies in multiple-choice based tests can inflate scores. In my test, the ‘I don’t know’ option was added in order to decrease guessing effects and to

increase test reliability (Pellicer-Sanchez & Schmitt, 2010; Stewart, 2014; Zhang, 2013). Item analysis could have revaled these effects even more clearly.

I acknowledged these limitations of multiple-choice testing by investigating the pooled test-data, and by splitting the test scores up into different analytical tiers.

However, these problems were obvious with only a few students; most participants were interested and worked carefully throughout all three sessions.

Other limitations of my study concern the interviews. Interviewing people about the fine details distinguishing the effect and use of different reading conditions was not as revealing as I had anticipated. It seems that many of the involved learning processes are subconscious. This became clear only when I was examining the data. It must be assumed that instead of using interviews, a more effective method might have been to use online think-aloud reports (e.g. Bowles, 2004; Leow, 2001). These might have provided better insight into the scale of word knowledge, so that the quantitative findings would have been complemented in a better way. This was considered in the original design of the research set-up, but was dismissed as it would have been too laborious to be used in this authentic classroom research. Unfortunately, the involved teachers, who were very generous with their time, or the researcher, did not have the capacities to facilitate the training needed for working successfully with such an intricate method. Moreover, I had also wanted to ask the interviewees about specific facets of TIE reading/learning and to react spontaneously to whatever relevant statements might emerge during the interviews. This would not have been possible with think-aloud protocols. In the end, this methodological decision was a trade-off between using an online processing measure which might have yielded rich natural data, and being able to actively interact with the participants.

In recent related studies, increasingly other, technologically more accurate measures of cognitive processing, such as eye-tracking, have been used. Godfroid et al. (2013) used eye-tracking as an online form of attention measurement. Eye-tracking measures attention through gauging eye fixations and eye movements (saccades). This technology makes it possible to “study the dynamics of cognitive processes online with a great amount of detail. A second benefit of eye-tracking is that it does not involve a secondary task and, therefore, does not carry a risk of altering the very process that it is intended to measure” (p. 509). Consequently, it is used increasingly in recent related studies (Choi, 2016; Godfroid & Schmidtke, 2013; Pellicer-Sanchez, 2012).

The other point of interest, TIE-related learner behaviour, however, could well be explored with interviews. Here, leading the interviews more professionally (explaining and interfering less, better prompting, questioning so that interesting threads would be pursued) would have

yielded more insightful results. Generally, think-aloud protocols might have revealed more of the learning processes, however, they were thought to be too difficult to carry out within my research design.

Finally, it needs to be acknowledged that my study was conducted in a specific learning context and that this specificity limits the generalizability of my findings. While one of the strengths of this study lies in the fact that it investigated an under-researched population, the specific context the data were gathered in raises the question of whether the results are applicable for the general population of EFL learners. It seems likely that reading behaviour and specifically the use of TIEs would be different in different settings (age group, country of setting, L1, proficiency). Would, for instance, university students in their 20s studying advanced level EFL in France show the same type of reading behaviour? As I, for instance, found that proficiency is decisive for TIE use, these considerations are highly relevant here. It must be assumed that reading and learning approaches vary considerably from setting to setting, so that my findings cannot necessarily be regarded as being valid for learning situations beyond the one investigated in my study.

7.4. Future Research

Through my research on the effects of TIEs on vocabulary learning and the connected learner behaviour, other questions and areas worth investigating have emerged. I would like to suggest five areas that need to be researched in the field of language learning through reading enhanced texts.

First, I recommend that future research should investigate the effects of TIEs with respect to aspects that my study did not shed sufficient light on. One is the effect of TIE use on long-term retention of vocabulary. Unfortunately, my delayed post-test data did not provide this information. However, knowing how well the effects of TIEs last over time is an essential part of knowing how well their application actually makes vocabulary learning more effective. Another insufficiently explored aspect relates to the fact that simultaneous processing of both vocabulary learning and reading comprehension is likely to be a strain on learners' cognitive capacities. Therefore, the question remains whether some of the vocabulary knowledge was acquired at the expense of reading comprehension. Future research is needed that investigates whether there is necessaryly a trade-off between these two learning processes, and forthcoming research into the efficacy of TIE should make the effort to integrate a measure of reading comprehension. It would also be valuable to investigate the effects of TIEs in

connection with frequency of exposure, as frequency has been found to be an important factor in vocabulary learning. For reasons explained in the Methodology Chapter, it was not possible to include frequency as a variable in my study.

Secondly, my study demonstrates that test taking behaviour and test attitude can influence test results. TIE related studies in particular, and vocabulary learning research in general, should investigate these aspects more thoroughly, for instance by carrying out a more systematic qualitative analysis than was possible within the scope of my study. Such investigations must take into account, for instance test fatigue, test length, the ratio of targeted vs. non-targeted words in reading materials, and the nature of the test instructions.

Thirdly, my study confirmed the findings of previous research with regards to the fact that the efficacy of TIEs is as much determined by the teacher (externally, through task design) as by the learner (internally, through behaviour; for instance O'Donnell, 2012; Sharwood Smith, 1993). More research is needed to understand the nature of both teacher- and learner-behaviour in relation to TIE use. One specifically interesting question is in how far it is task-design or student behaviour that determines whether vocabulary in an assignment is handled incidentally or intentionally.

Forthly, my study explored the learning of single words. However, vocabulary is often learned in the form of lexical bundles (e.g. collocations, phrasal verbs). As this has implications for vocabulary teaching in foreign language classrooms, research should, and already has begun (Boers et al., 2016; Choi, 2016) to provide evidence of how TIEs can help learners to acquire lexical bundles.

Finally, the way TIEs are used has changed profoundly since I started working on this research project. While I conducted my testing sessions with the participants filling in their responses with pens on paper, nowadays learners are much more often exposed to new vocabulary while reading texts on computers screens. The latest technological developments offer new possibilities of presenting unknown vocabulary to readers (e.g. with links to pictures, videos, pronunciation help, and various (extra-lingual) information). To some extent glosses are provided throughout the internet. New TIE studies could therefore investigate the effect of computerized enhancements. New technology also offers new methodological possibilities for vocabulary research (e.g. by tracking clicks on glosses, eye-tracking movement). Research should investigate the impact of new technology on enhanced word learning and the challenges that come with it.

The recent publication of several TIE related studies (e.g. Boers et al., 2016; Jung, 2016; LaBrozzi, 2016; Peters, 2016) shows that this is a vibrant field of research with many unexplored questions. I hope that the work presented in my study contributes to furthering the

understanding of how textual enhancements can aid vocabulary learning and that it can be of benefit for foreign language learners and teachers.

BIBLIOGRAPHY

Abraham, L. B. (2008). Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. *Computer Assisted Language Learning*, 21 (3), 199–226.

Aitchison, J. (1994). *Words in the Mind. An Introduction to the Mental Lexicon*. Oxford: Blackwell.

Alanen, R. (1995). Input enhancement and rule presentation in second language acquisition. In R. Schmidt (Ed.), *Attention and Awareness in Second Language Acquisition* (259–299). Honolulu: University of Hawai'i Press.

Allen, P., Swain, M., Harley, B., & Cummins, J. (1990). Aspects of classroom treatment: Toward a more comprehensive view of second language education. In B. Harley, P. Allen, J. Cummins & M. Swain (Eds.). *The development of second language proficiency* (57-81). Cambridge: Cambridge University Press.

Allendorf, S. & Wode, H. (1981). Some overgeneralizations in the L1 acquisition of interrogative pronouns. *International Review of Applied Linguistics*, 19, 31-44.

Alloway, T. & Gathercole, S. (2005). *Working memory. Teaching Thinking and Creativity*. Retrieved from http://www.york.ac.uk/res/wml/AllowayTTC.pdf, accessed on 14/2, 2014.

Alptekin, C., & Ercetin, G. (2011). The effects of working memory capacity and content familiarity on literal and inferential comprehension in L2 reading. *TESOL Quarterly*, 45, 235–266.

Alqahtani, M. (2015). The importance of vocabulary in language learning and how to be taught. *International Journal of Teaching and Education*, 3 (3), 21-34.

Al-Seghayer, K. (2001). The Effect of Multimedia Annotation Modes On L2 Vocabulary Acquisition: A Comparative Study. *Language Learning & Technology*, 5 (1), 202-232.

Alter, A. L. & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, 13 (3), 219–235.

Anderson, R. & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and Teaching: Research Reviews* (77-115). Newark, DE: International Reading Association.

Angouri, J. (2010). Quantitative, Qualitative or Both? Combining Methods in Linguistic Research. In L. Litosseliti (Ed.), *Research Methods in Linguistics* (29-45). London, New York: Continuum.

Azari, F. (2012). Review of Effects of Textual Glosses on Incidental Vocabulary Learning. *International Journal of Innovative Ideas*, 12 (2), 13-24.

Azari, F., FaizSathi A., Hoon, T. B. (2012). Effects of Textual Glosses On Reading Comprehension Of Low Proficiency EFL Postgraduates. *The English Teacher,* 41 (1), 42-55.

Baddeley, A.D. (1978). The trouble with levels: A re-examination of Craik and Lockhart's framework for memory research. *Psychological Review*, 85, 139–152.

Baddeley, A. D. (1997). *Human Memory. Theory and Practice*. Hove, East Sussex: Psychology Press.

Bahns, J. (1993). Lexical collocations: a contrastive view. *ELT Journal*, 47 (1), 56–63.

Balan, P. (1989). Improving instructional print materials through text design. *Performance and Instruction*, 28 (7), 13–18.

Barcroft, J. (2002). Semantic and structural elaboration in L2 lexical acquisition. *Language Learning*, 52, 323–363.

Barcroft, J. (2003). Distinctiveness and Bidirectional Effects in Input Enhancement for Vocabulary Learning. *Applied Language Learning*, 3 (2), 47-73.

Barcroft, J. (2004). Effects of sentence writing in L2 lexical acquisition. *Second Language Research*, 20, 303–334.

Barcroft, J. (2009b). Effects of Synonym Generation on Incidental and Intentional L2 Vocabulary Learning During Reading. *TESOL Quarterly,* 43 (1), 79-103.

Bell, F. L. & LeBlanc, L. B. (2000). The language of glosses in L2 reading on computer: Learners' preferences. *Hispania*, 83 (2), 274–285.

Bishop, H. (2004). The effect of typographic salience on the look up and comprehension of unknown formulaic sequences. In N. Schmitt (Ed.) *Formulaic Sequences*. *Language Learning & Language Teaching* (227-248). Amsterdam: John Benjamins.

Bland, S. K., Noblitt, J. S., Armington, S., Gay, G. (1990). The Naive Lexical Hypothesis: Evidence from Computer-Assisted Language Learning. *The Modern Language Journal*, 74 (4), 440-450.

Block, D. (1996). A window on the classroom: classroom events viewed from different angles. In K. M. Bailey & D. Nunan (Eds.), *Voices from the Language Classroom* (168-196). Cambridge: Cambridge University Press.

Bloom, P. (2000). *How Children Learn the Meanings of Words*. Harvard, MA: MIT Press.

Boers, F., Demecheleer, M., He, L., Deconinck, J., Stengersband, H., Eyckmans, J. (2016). Typographic enhancement of multiword units in second language text. *International Journal of Applied Linguistics*, 26 (1).

Bowles, M. A. (2004). L2 glossing: To CALL or not to CALL. *Hispania*, 87 (3), 541–552.

Brantmeier, C., Van Bishop, T., Yu, X., Anderson, B. (2012). Readings on L2 Reading: Publications in Other Venues 2011-2012. *Reading in a Foreign Language*, 24 (2), 256–272.

Brenner, M. E. (2006). Interviewing in Educational Research. In J. L. Green, G. Camilli & P. B. Elmore (Eds.), *Handbook of Complementary Methods in Education Research* (357-370). Mahwah, NJ: Lawrence Erlbaum.

Brown, R. (2000). Extensive reading in action. *Studies in English Language and Literature*, 41, 79–123.

Brown, R., Waring, R., Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, 20 (2), 136-163.

Bruton, A. (2009). The Vocabulary Knowledge Scale: A Critical Analysis. *Language Assessment Quarterly,* 6 (4), 288-297.

Bruton, A., Lopez, M. G., Mesa, R.E. (2011). Incidental L2 Vocabulary Learning: An Impracticable Term? *TESOL Quarterly*, 45 (4), 759-768.

Bryman, A. (2006). Integrating quantitative and qualitative research: how is it done? *Qualitative Research*, 6 (1), 97-113.

Bryman, A. (2008). *Social research methods* (3rd ed.). Oxford: Oxford University Press.

Bryman, A. (2012). *Social research methods* (4th ed.). Oxford: Oxford University Press.

Carrier, M. & Pashler, H. (1992). The influence of retrieval on retention. *Memory and Cognition*, 20 (6), 633-642.

Carroll, S. (1999). Putting 'input' in its proper place. *Second Language Research*, 15 (4), 337-388.

Carver, C. P. (1990). *Reading Rate*. San Diego, CA: Academic Press.

Charmaz, K. (2000). Constructivist and objectivist grounded theory. In N. K. Denzin & Y. Lincoln (Eds.), *Handbook of Qualitative Research* (2nd ed., 509–535). Thousand Oaks, CA: Sage.

Cheng, Y. & Good, R. (2009). L1 glosses: Effects on EFL learners' reading comprehension and vocabulary retention. *Reading in a Foreign Language*, 21 (2), 119-142.

Cho, M. Y. (2010). The effects of input enhancement and written recall on noticing and acquisition. *Innovation in Language Learning and Teaching*, 4 (1), 71–87.

Choi, S. (2016). Processing and learning of enhanced English collocations: An eye movement study. Language Teaching Research, 20 (1), 1-24.

Clark, E. (1993). *The Lexicon in Acquisition*. Cambridge: Cambridge University Press.

Clarke, M. (1980). The short circuit hypothesis of ESL reading—or when language competence interferes with reading performance. *Modern Language Journal*, 64, 203-209.

Coady, J. (1997). L2 Vocabulary Acquisition through Extensive Reading. In J. Coady & T. Huckin (Eds.), Second Language Vocabulary Acquisition (225-238). Cambridge: Cambridge University Press.

Cobb, T. (1999a). Breadth and depth of vocabulary acquisition with hands-on concordancing. *Computer Assisted Language Learning*, 12, 345 - 360.

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning and Technology,* 11, 38-63.

Cohen, A.D. (2007): Coming to terms with language learner strategies: Surveying the Experts. In A. D. Cohen & E. Macaro. *Language Learner Strategies: Thirty Years of Research and Practice* (29-45). Oxford: Oxford University Press.

Cohen, A. D. (2011). L2 learner strategies. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning*, Vol. II - Part V. Methods and instruction in second language teaching (681-698). Abingdon, England: Routledge.

Cohen, A. D. & Macaro E. (2007). *Language Learner Strategies: Thirty Years of Research and Practice*. Oxford: Oxford University Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. London: Routledge.

Cohen, J., Cohen, P. & Stephen, G. (2003). *Applied Multiple Regression/Correlation Analysis For The Behavioral Sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Coolican, H. (2009). *Research methods and statistics in psychology*. Hodder Education.

Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics,* 5, 161-169.

Coxhead, A. (2000). A new "Academic Word List." *TESOL Quarterly*, 34 (2), 213-238.

Craik, F. I. M. (2002). Levels of processing: Past, present ... and future? *Memory*, 10 (5/6), 305-318.

Craik, F. I. M. & Lockhart, R. S. (1972). Levels of processing: a framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671-684.

Craik, F. I. M. & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology General,* 104 (3), 268-294.

Creswell, J. W. (2014) Research Design. Qualitative, Quantitative, and Mixed Methods Approaches (4th ed.). New York: Sage Publications.

Crystal, D. (2003). *English as a global language*. Cambridge: Cambridge University Press.

Dakin, J. (1973). *The Language Laboratory and Language Learning*. London: Longman.

Damskier, J. & Weidick, K. (2003). *Crossroads 1*. Århus: Systime.

Davis, J. N. (1989). Facilitating effects of marginal glosses on foreign language reading. *The Modern Language Journal*, 73 (1), 41-48.

Day, R., Omura, C., & Hiramatsu, M. (1991). Incidental EFL Vocabulary Learning and Reading. *Reading in a Foreign Language*, 7 (2), 541-551.

DeKeyser, R. (2001). Automaticity and automatization. In P. Robinson (Ed.), *Cognition and second language instruction* (125-151). Cambridge: Cambridge University Press.

DeKeyser, R. (2003). Implicit and explicit learning. In C. J. Doughty & M. H. Long (Eds.) *The Handbook of Second Language Acquisition* (313-348). Oxford: Blackwell Publishing.

Dörnyei, Z. (2007). *Research Methods in Applied Linguistics*. Oxford: Oxford University Press.

Dörnyei, Z. (2009). *The psychology of second language acquisition*. Oxford: University Press.

Dörnyei, Z. & Kormos, J. (2000). The role of individual and social variables in task performance. *Language Teaching Research*, 4, 275-300.

Doughty, C. (1991). Second Language Instruction Does Make A Difference - Evidence from an Empirical Study of SL Relativization. *Studies in Second Language Acquisition*, 13 (4), 431-469.

Doughty, C. (2004). Effects of instruction on learning a second language: A critique of instructed SLA research. In B. VanPatten, J. Williams, S. Rott, and M. Overstreet (Eds) *Form-meaning Connections in Second Language Acquisition* (181–202). Mahwah, NJ: Lawrence Erlbaum.

Doughty, C. & Long, M. (Eds.) (2003). *The handbook of second language acquisition*. Oxford: Blackwell Publishing.

Dupuy, B., & Krashen, S. (1993). Incidental vocabulary acquisition in French as a foreign language. *Applied Language Learning*, 4, 55–63.

Eckerth, J. (2009). Negotiated interaction in the L2 classroom. *Language Teaching,* 42 (1), 109-130.

Eckerth, J. & Tavakoli, P. (2012). The effects of word exposure frequency and elaboration of word processing on incidental L2 vocabulary acquisition through reading. *Language Teaching Research*, 16 (2), 227-252.

Elley, W. B., & Mangubhai, F. (1983). The impact of reading on second language learning. *Reading Research Quarterly*, 19 (1), 53-68.

Ellis, N. (1994). Consciousness in second language learning: Psychological perspectives on the role of conscious processes in vocabulary acquisition. *Consciousness in Second Language Learning*, 11, 37–57.

Ellis, N. (1994b). Implicit and Explicit Learning – An Overview. Introduction. In N. C. Ellis (Ed.) Implicit and explicit learning of languages (1-31). San Diego, CA: Academic Press.

Ellis, N. (1994c). Vocabulary acquisition: The implicit ins and outs of explicit cognitive mediation. In N. C. Ellis (Ed.) *Implicit and explicit learning of languages* (211–282). San Diego, CA: Academic Press.

Ellis, N. (1995b). The psychology of foreign language vocabulary acquisition: Implications for CALL. *Computer Assisted Language Learning*, 8, 103–128.

Ellis, N. (2001). Memory for language. In P. Robinson (Ed.) *Cognition and second language instruction* (33–68). New York: Cambridge University Press.

Ellis, N. (2004). The processes of second language acquisition. In B. VanPatten, J. Williams, S. Rott, & M. Overstreet (Eds.). *Form-meaning connections in second language acquisition* (49-76). Mahwah, NJ. : Erlbaum.

Ellis, N. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition*, 27, 305–352.

Ellis, N. (2006). Selective Attention and Transfer Phenomena in L2 Acquisition: Contingency, Cue Competition, Salience, Interference, Overshadowing, Blocking, and Perceptual Learning. *Applied Linguistics*, 27 (2), 164-194.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Ellis, R., Tanaka, Y. & Yamakazi, A. (1994). Classroom interaction, comprehension and L2 vocabulary acquisition. *Language Learning,* 44, 449–491.

Engberg-Pedersen, J., Grønvold, M., & Ohland-Andersen, H. (2004). *Contexts*. Copenhagen: Gyldendal.

Erler, L. & Finkbeiner, C. (2007). A review of reading strategies: focus on the impact of first language. In J. Cohen, D. Andrew & E. Macaro (Eds.). *Language Learner Strategies* (187-206). Oxford: Oxford University Press.

Eskey, D. E. (2005). Reading in a Second Language. In E. Hinkel (Ed.) *Handbook of Research in Second Language Teaching and Learning* (563-579). Mahwah, New Jersey, Erlbaum.

Fahimipour, N. & Hashemian, M. (2013). The effect of lexical glossing types on reading and listening skills of Iranian EFL learners. *Journal of Language Teaching and Research*, 4 (1), 192-199.

Færch, C., Haastrup, K. and Phillipson, R. (1984). *Learner Language and Language Learning*. Copenhagen: Gyldendals Sprogbibliotek.

Ferguson, C. (1971). Absence of copula and the notion of simplicity: a study of normal speech, baby talk, foreigner talk and pidgins. In D. Hymes (Ed.), *Pidginization and Creolization of Languages* (141-150). Cambridge: Cambridge University Press.

Field, A. (2005). *Discovering statistics with SPSS*. London: Sage.

Finkbeiner, C., Knierim, M., Smasal, M., & Ludwig, P. H. (2005). Self-regulated cooperative EFL reading tasks: students' strategy use. *Language Awareness*, 21 (1/2), 57-83.

Folse, K. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, 40 (2), 273-293.

Folse, K. S. (2012). Myths About Second Language Vocabulary Learning. In C. A. Chapelle (Ed.). *The Encyclopedia of Applied Linguistics*. Boston: Wiley-Blackwell.

Fraser, C. (1999). Lexical processing strategy use and vocabulary learning through reading. *Studies in Second Language Acquisition*, 21, 225-241.

Fuente, de la M. J. (2002). Negotiation and oral Acquisition of L2 Vocabulary. *Studies in Second Language Acquisition,* 24 (01), 81-112.

Fuente, de la M. J. (2006). Classroom L2 vocabulary acquisition: Investigating the role of pedagogical tasks and form-focussed instruction. *Language Teaching Research*, 10, 263-295.

Fukkink, R. G., Blok, H. & de Glopper, K. (2001). Deriving word meaning from written context: a multicomponential skill. *Language Leaning*, 51 (3), 477-496.

Gascoigne, C. (2006). Explicit input enhancement: Effects on target and non-target aspects of second language acquisition. *Foreign Language Annals*, 39 (4), 551-564.

Gass, S. (1988). Integrating research areas: a framework for second language studies. *Applied Linguistics,* 9 (2), 198-217.

Gass, S. (1997). *Input, Interaction, and the Second Language Learner*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gass, S. (1999). Incidental vocabulary learning. *Studies in Second Language Acquisition*, 21 (2), 319-333.

Gass, S. (2010). Experimental Research. In B. Paltridge & A. Phakiti (Eds.) *Research Methods in Applied Linguistics* (7-21). London & New York: Continuum.

Gass, S. and Selinker, L. (2008). *Second Language Acquisition*. New York: Routledge.

Gee, J. (2000). Discourse and sociocultural studies in reading. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research*, Vol III (195-208). Mahwah, NJ: Lawrence Erlbaum Associates

Gettrich-Ludgate, B. & Tovar, D. (1987). Authentic Text Types and Corresponding Strategies: A List for the Foreign Language Instructor? *Die Unterrichtspraxis,* 20, 80-94.

Gettys, S., Imhof, L. A., Kautz, J. O. (2001). Computer-Assisted Reading: The Effect of Glossing Format on Comprehension and Vocabulary Retention. *Foreign Language Annals*, 34 (2), 91-99.

Glesne, C., & Peshkin, P. (1992). *Becoming qualitative researchers: An introduction*. New York, NY: Longman.

Godfroid, A., Boers, F., Housen, A. (2013). An Eye For Words - Gauging the Role of Attention in Incidental L2 Vocabulary Acquisition by Means of Eye-Tracking. *Studies in Second Language Acquisition*, 35, 483-517.

Godfroid, A. & Schmidtke, J. (2013). *What Do Eye Movements Tell Us About Awareness? A Triangulation of Eye-Movement Data, Verbal Reports, and Vocabulary Learning Scores*. In J. M. Bergsleithner, S. N. Frota, & J. K. Yoshioka (Eds.), Noticing and second language acquisition: Studies in Honour of Richard Smith. Honolulu: University of Hawai'I, National Foreign Language Resource Center, 183-205.

Golafshani, N. (2003). Understanding Reliability and Validity in Qualitative Research. *The Qualitative Report,* 8 (4), 597-607.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York, NY: Cambridge University Press.

Greene, J. C., Caracelli, V. J. & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, 11, 255-274.

Griffin, G. F. (1992). *Aspects of the psychology of second language vocabulary list learning*. Unpublished doctoral dissertation. Dept. of Psychology, University of Warwick.

Grissom, R. J. & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate observations*. New York: Routledge.

Gu, Y. (2012). Vocabulary Learning Strategies. In C. A. Chapelle (Ed.). *The Encyclopedia of Applied Linguistics*. Boston. Wiley-Blackwell.

Gu, Y. & Johnson, R. K. (1996) Vocabulary learning strategies and language learning outcomes. *Language Learning*, 46 (4), 643-679.

Gyllstad, H., Vilkaite, L. & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL International Journal of Applied Linguistics*. 166 (2), 276-303.

Haastrup, K. (1991). *Lexical Inferencing Procedures or Talking about Words*. Tübingen, Germany: Gunter Narr.

Haastrup, K (2008). Lexical inferencing procedures in two languages. In D. Albrechtsen, K. Haastrup & B. Henriksen. *Vocabulary and Writing in a First and Second Language. Processes and Development*. Basingstoke: Palgrave Macmillan.

Haastrup, K. & Henriksen, B. (1998). Vocabulary Acquisition: from partial to precise understanding. In K. Haastrup, Å. Viberg (Eds.) *Perspectives on Lexical Acquisition in a Second Language* (97-124). Lund University Press.

Haastrup, K. & Henriksen, B. (2000a). Vocabulary acquisition: acquiring depth of knowledge

through network building. *International Journal of Applied Linguistics*, 10, 221-239.

Haastrup, K. & Henriksen, B. (2000b). *Lexical inferencing in relation to comprehension and vocabulary acquisition*. Unpublished Manuscript.

Han, Z. H., Park, E. S. & Combs, C. (2008). Textual enhancement of input: Issues and possibilities. *Applied Linguistics*, 29 (4), 597–618.

Harmer, J. (2001). *The Practice of English Language Teaching* (3rd Ed.). Harlow: Pearson Education Limited.

Harwell, M. R. & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*. 71 (1), 105-131.

Hatch, EM and Brown, C (1995). *Vocabulary, Semantics, and Language Education*. Cambridge: Cambridge University Press.

Heiman, G. (2014). *Behavioral Sciences STAT 2*. Cengage Learning.

Henriksen. (1999). Three Dimensions of Vocabulary Development. *Studies in Second Language Acquisition*, 21, 303-317.

Herman, P., Anderson, R., Pearson, D., & Nagy, W. (1987). Incidental Acquisition of Word Meaning from Expositions with Varied Text Features. *Reading Research Quarterly*, 22 (3), 263-282.

Holley, F. M. & King, J. K. (1971). Vocabulary glosses in foreign language reading materials. *Language Learning*, 21 (2), 213–219. Wiley Online Library.

Holliday, A. (2010). Analysing Qualitative Data. In B. Paltridge & A. Phakiti (Eds.), *Research Methods in Applied Linguistics* (98-110). London & New York: Continuum.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, 11 (2), 207-223.

Howell, D. C. (2011). *Fundamental statistics for the behavioural sciences*. Cengage Learning. Wadsworth Publishing.

Howell, D. (2013, 8th edition). *Fundamental statistics for the behavioral sciences*. Cengage Learning. Wadsworth Publishing.

Hu, M., & Nation, P. (2000). Unknown Vocabulary Density and Reading Comprehension *Reading in a Foreign Language,* 13 (1), 403-430.

Huang, S., Eslami, Z., Willson, V. (2012). The Effects of Task Involvement Load on L2 Incidental Vocabulary Learning: A Meta-Analytic Study. *The Modern Language Journal*, 96 (4), 544-557.

Huckin, T. and Bloch, J. (1993). Strategies for inferring word meanings in context: a cognitive model. In T. Huckin, M. Haynes & J. Coady (Eds.). *Second Language Reading and Vocabulary Learning* (153-179). Norwood, NJ: Ablex.

Huckin, T and Coady, J (1999). Incidental vocabulary acquisition in a second language. A review. *Studies in Second Language Acquisition,* 21 (2), 181-193.

Hulstijn, J. (1992). Retention of inferred and given word meanings: experiments in incidental vocabulary learning. In. P. Arnaud & H. Bejoint (Eds.). *Vocabulary and Applied Linguistics* (113-125). London: MacMillan.

Hulstijn, J. H. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.). *Cognition and second language instruction* (258-287). Cambridge: University Press.

Hulstijn, J. (2005). Theoretical and empirical issues in the study of implicit and explicit second-language learning: Introduction. *Studies in Second Language Acquisition*, 27, 129-140.

Hulstijn, J. (2012). Incidental Learning in Second Language Acquisition. In C. A. Chapelle (Ed.). *The Encyclopedia of Applied Linguistics*. Boston: Wiley-Blackwell.

Hulstijn, J. H., Hollander, M. & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal*, 80 (3), 327–339.

Hulstijn, J. H. & Laufer, B. (2001). Some empirical evidence for the Involvement Load Hypothesis in Vocabulary Acquisition. *Language Learning,* 51 (3), 539-558.

Iser, Wolfgang (1972). The reading process: a phenomenological approach. *New Literary History*, 3 (2), 279-299.

Izumi S. (2002). Output, input enhancement, and the noticing hypothesis. *Studies in Second Language Acquisition*, 24 (4), 541-577.

Jacobs, G. M., Dufon, P., Hong, F. C. (1994). L1 and L2 vocabulary glosses in L2 reading passages: Their effectiveness for increasing comprehension and vocabulary knowledge. *Journal of Research in Reading*, 17 (1), 19-28.

Jensen, L. S. (2005). *Vocabulary Knowledge and Listening Comprehension in English as a Foreign Language: An Empirical Study Employing Data Elicited From Danish EFL Learners.* Copenhagen: Samfundslitteratur.

Jiang, N (2000). Lexical representation and development in a second language. *Applied Linguistics*, 21 (1), 47-77.

Jiménez, R. T., García, G. E., & Pearson, P. D**.** (1996). The reading strategies of bilingual Latina/o students who are successful English readers: Opportunities and obstacles. *Reading Research Quarterly*, 31 (1), 90–112.

Joe, A. (1995). Text based tasks and incidental vocabulary learning. *Second Language Research,* 11 (2), 149-158.

Joe, A. (1998). What effects do text-based tasks promoting generation have on incidental vocabulary acquisition? *Applied Linguistics,* 19 (3), 357-377.

Jooneghani, R. B., Jozani, N. E., Mei, L. L. (2012). Language Learning Strategy: Teachable Or Learnable? *International Journal of Academic Research*, 4 (6), 75-78.

Joppe, M. (2000). The Research Process. *The Quantitative Report Journal*. 8 (4), 597-607. Retrieved from http://www.nova.edu/ssss/QR/QR8-4/golafshani.pdf

Jourdenais, R. (1998). *The effects of textual enhancement on the acquisition of the Spanish preterit and imperfect*. Doctoral dissertation. Retrieved from ProQuest Dissertations and Theses databases. (UMI No. 9924386)

Jourdenais, R. (2001). Cognition, instruction, and protocol analysis. In P. Robinson (Ed.). *Cognition and Second Language instruction* (354-375). Cambridge: Cambridge University Press.

Jourdenais, R., Ota, M., Stauffer, S., Boyson, B. & Doughty, C. (1995). Does textual enhancement promote noticing? A think-aloud protocol analysis. In R. Schmidt (Ed.). *Attention and awareness in foreign language learning* (183-216). Honolulu, HI: University of Hawaii.

Juffs, A. & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching*. 44 (2), 137-166.

Jung, J. (2016). Effects of glosses on learning of L2 grammar and vocabulary. *Language Teaching Research*, 20 (1), 92-112.

Keating, G. D. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research,* 12 (3), 365-386.

Keats, D. M. (2000). *Interviewing. A practical guide for students and professionals*. Buckingham. Philadelphia: Open University Press.

Kim, Y. (2006). Effects of Input Elaboration on Vocabulary Acquisition through Reading by Korean Learners of English as a Foreign Language. *TESOL Quarterly*, 40 (2), 341-373.

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning,* 58 (2), 285-325.

Klapper, J. (2006). *Understanding and Developing Good Practice: Teaching Languages in Higher Education*. London: CILT.

Ko, M. H. (2012). Glossing and Second Language Vocabulary Learning. *TESOL Quarterly*, 46 (1), 56–79.

Koda, K. (1996). L2 word recognition research: A critical review. *The Modern Language Journal*, 80 (4), 450-406.

Koda, K. (2005). *Insights into Second Language Reading: A Cross-Linguistic Approach*. Cambridge: Cambridge University Press.

Kojic-Sabo, I. & Lightbown, P. (1999). Students' approaches to vocabulary learning and their relationship to success. *The Modern Language Journal*, 83 (2), 176-192.

Kost, C.R., Foss, P., Lenzini Jr, J. J. (1999) Textual and pictorial glosses: Effectiveness on incidental vocabulary growth when reading in a foreign language. *Foreign Language Annals*, 32 (1), 89-97.

Krashen, S. (1989b). We acquire vocabulary and spelling by reading: Additional evidence for the Input Hypothesis. *The Modern Language Journal*, 73, 440-464.

Krashen, S. (1993). *The power of reading*. Englewood, NJ: Libraries Unlimited.

LaBrozzi, R. M. (2016). The effects of textual enhancement type on L2 form recognition and reading comprehension in Spanish. *Language Teaching Research*, 20 (1), 75-91.

Lahav, J. (1996). Vocabulary Acquisition Through extensive Reading: 'Incidental' and Controlled Acquisition Compared. MA Dissertation, University of Haifa.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. London: Routledge.

Laufer, B. (1988). The concept of 'synforms' (similar lexical forms) in vocabulary acquisition. *Language and Education*, 2 (2), 113–132.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.). *Special language: From humans thinking to thinking machines* (316-323). Clevedon: Multilingual Matters.

Laufer, B. (1997b). What's in a word that makes it hard or easy? Intralexical factors affecting the difficulty of vocabulary acquisition. In N. Schmitt and M. McCarthy (Eds.). *Vocabulary: Description, Acquisition, and Pedagogy* (140-155). Cambridge: Cambridge University Press.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: same or different? *Applied Linguistics*, 19 (2), 255-271.

Laufer, B. (2001). Quantitative evaluation of vocabulary: How it can be done and what it is good for. In C. Elder, K. Hill, A. Brown, N. Iwashita, L. Grove, T. Lumley, & T. McNamara (Eds.). *Experimenting with Uncertainty* (241-250). Cambridge: Cambridge University Press.

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review*, 59 (4), 567-587.

Laufer, B. (2005). Focus on form in second language vocabulary acquisition. In S. H. Foster-Cohen, M. P. Garcia-Mayo and J. Cenoz (Eds.). *EUROSLA Yearbook 5* (223-250). Amsterdam: Benjamins.

Laufer, B. (2012). Vocabulary and writing. In C. A. Chapelle (Ed.). *The Encyclopedia of Applied Linguistics*. Boston: Wiley-Blackwell.

Laufer. B. (2012). Word difficulty. In C. A. Chapelle (Ed.). *The Encyclopedia of Applied Linguistics*. Boston: Wiley-Blackwell.

Laufer, B., Elder, C., Hill, K. & Congdon, P. (2004). Size and strength: do we need both to measure vocabulary knowledge? *Language Testing*, 21 (2), 202–226.

Laufer, B. & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, 54 (3), 399–436.

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, 22 (1), 1-26.

Laufer, B., Meara, P., and Nation, I. S. P. (2005). Ten best ideas for teaching vocabulary. *The Language Teacher (Japan),* 29 (5), 3-6.

Laufer, B. & Nation, P. (1995). Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics*, 16, 307-322.

Laufer, B. & Nation, I. S. P. (2011) Vocabulary. In S. Gass & A. Mackey. *The Routledge Handbook of Second Language Acquisition* (163-176). London and New York: Routledge.

Laufer, B. and Paribakht, T. S. (1998). The relationship between passive and active vocabularies: effects of language learning context. *Language Learning*, 48 (3), 365–391.

Laufer, B. & Ravenhorst-Kalovski. G. C. (2010). Lexical threshold revisited: lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22 (1), 15-30.

Laufer, B. & Shmueli, K. (1997). Memorizing new words: does teaching have anything to do with it? *RELC Journal*, 28 (1), 89-108.

Laufer, B. & Sim, D. D. (1985). Taking the easy way out: non-use and misuse of clues in EFL reading. *English Teaching Forum*, 23, 7-20.

Laufer, B. and Yano, Y. (2001). Understanding unfamiliar words in a text: Do L2 learners understand how much they don't understand? *Reading in a Foreign Language,* 13 (2), 549-566.

Lawson, M. J. and Hogben, D. (1996). The vocabulary learning strategies of foreign-language students. *Language Learning*, 46 (1), 101-135.

Lazaraton, A. (2000). Current trends in research methodology and statistics in applied linguistics, *TESOL Quarterly*, 34 (1), 175-181.

Lazaraton, A. (2005). Quantitative Research Methods. In E. Hinkel (Ed.). *Handbook of research in Second Language Teaching and Learning* (209-221). Mahwah, NJ: Lawrence Erlbaum.

Lee, S. K. (2007). Effects of textual enhancement and topic familiarity on Korean EFL students' reading comprehension and learning of passive form. *Language Learning*, 57 (1), 87-118.

Lee, S. & Huang, H. (2008). Visual input enhancement and grammar learning. *Studies in Second Language Acquisition*, 30 (3), 307–331.

Leeman, J., Arteagoitia, I., Fridman, B., & Doughty, C. (1995). Integrating attention to form with meaning: Focus on form in content-based Spanish instruction. In R. Schmidt (Ed.). *Attention and awareness in foreign language learning* (217–258). Honolulu: University of Hawaii Press.

Leow, R. P. (1993). To simplify or not to simplify: A look at intake. *Studies in Second Language Acquisition*, 15, 333-355.

Leow, R. P. (1995). Modality and intake in SLA. *Studies in Second Language Acquisition*, 17, 79 – 89.

Leow, R. P. (2000). A study in the role of awareness in foreign language behaviour: aware versus unaware learners. *Studies in Second Language Acquisition*, 22, 557-584.

Leow, R. P. (2001). Do learners notice enhanced forms while interacting with the L2? An online and offline study of the role of written input enhancement in L2 reading. *Hispania*, 84 (3), 496–509.

Leow, R. P., Egi, T., Nuevo, A. M. & Tsai, Y.-C. (2003). The roles of textual enhancement and type of linguistic item in adult L2 learners' comprehension and intake. *Applied Language Learning*, 13, 93–108.

Leow, R. P. (2015). *Explicit Learning in the L2-Classroom: A Student-Centered Approach*. London: Routledge.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Lewis, M. (1993). *The Lexical Approach. The state of ELT and the way forward*. Hove: Language Teaching Publications.

Lewis, M. (1997). *Implementing the Lexical Approach: Putting Theory into Practice*. Hove, England: Language Teaching Publications.

Li, X. (1988). Effects of contextual cues on inferring and remembering meanings of new words. *Applied Linguistics*, 9 (4), 402-413.

Litosseliti, L. (2010). *Research Methods in Linguistics*. London: Continuum.

Liu, & Nation. (1985). Factors affecting guessing vocabulary in context. *RELC Journal*, 16, 33-42.

Loewen, S. & Gass, S. (2009). The use of statistics in L2 acquisition research. *Language Teaching, 42* (2), 181-196.

Lomicka, L. (1998). To gloss or not to gloss: An investigation of reading comprehension online. *Language Learning & Technology*, 1 (2), 41-50.

Long, M. H. (2000). Second language acquisition theories. In M. Byram (Ed.). *Encyclopedia of language teaching* (527-34). London: Routledge.

Long, J. S. & Freese, J. (2006). *Regression models for categorical dependent variables using Stata*. Stata press.

Long, M. H. & Robinson, P. (1998). Focus on form: Theory, research, and practice. In C. J. Doughty & J. Williams (Eds.). *Focus on form in second language acquisition* (15 – 41). New York: Cambridge University Press.

López, M. G. (2012). The Author's Reply. *TESOL Quarterly*, 46 (4), 817-819.

Lowie, W. & Seton, B. (2012). *Essential statistics for applied linguistics*. Palgrave Macmillan.

Lyman-Hager, M., & Davis, J.F. (1996). The case for computer-mediated reading: Une vie de boy. *The French Review*, 69, 775–790.

Macalister, J. (2011). Today's teaching, tomorrow's text: exploring the teaching of reading. *ELT Journal,* 65 (2), 161-169.

Mackey, A & Philp, J. (1998) Conversational interaction and second language development: recasts, responses, and red herrings. *Modern Language Journal*, 82, 338-356.

Mackey, A., Gass, S.M., McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition*. 22 (4), 471-497.

Mackey, A., Philp, J., Egi, T., Fujii, A., & Tatsumi, T. (2002). Individual differences in working memory, noticing of interactional feedback and L2 development. In P. Robinson (Ed.). *Individual differences and instructed language learning* (181–209). New York: Benjamins.

Mackey, A., & Gass, S. M. (2005). *Second Language Research*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Maggetti, M., Radaelli, C. M. & Gilardi, F. (2012). *Designing research in the social sciences*. Sage: London.

Maggetti. M., Gilardi, F., Radaelli, C. M. (2013). *Designing Research in the Social Sciences*. Sage: London.

McDonald, J. F. & Moffitt, R. A. (1980). The uses of Tobit analysis. *The review of economics and statistics*, 318–321.

McNeill, A. (1996). Vocabulary knowledge profiles: Evidence from Chinese-speaking ESL teachers. *Hong Kong Journal of Applied Linguistics*, 1, 39-63.

Meara, P. (1990). Some notes on the Eurocentres vocabulary tests. In J. Tommola (Ed.). *Foreign language comprehension and production* (103-113). Turku: AFinLa Yearbook.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.). *Performance and competence in second language acquisition* (35-52). Cambridge: University Press.

Meara, P. (1997). Towards a new approach to modelling vocabulary acquisition. In Schmitt, N. and McCarthy, M. (eds.), *Vocabulary: Description, Acquisition, and Pedagogy* (109 -121) Cambridge: Cambridge University Press.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, 4, 142-154.

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt and M. McCarthy (Eds.). *Vocabulary: Description, Acquisition, and Pedagogy.* Cambridge: Cambridge University Press.

Miles, M. B. & Huberman M. A. (1994, 2nd ed.). *Qualitative Data Analysis: An Expanded Sourcebook*. London: Sage Publications.

Milton, J. (2008). Vocabulary uptake from informal learning tasks. *Language Learning Journal* 36 (2), 227-237.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual Matters.

Min, H. (2008). EFL Vocabulary Acquisition and Retention: Reading Plus Vocabulary Enhancement Activities and Narrow Reading. *Language Learning,* 58 (1), 73-115.

Miyake, A. & Friedman, N. (1998). Individual differences in second language proficiency: working memory as language aptitude. In A. Healy & L. Bourne Jr. (Eds.). *Foreign Language Learning: Studies on Training and Retention* (339-364). Mahwah, NJ: Lawrence Erlbaum Associates.

Miyake, A. & Shah, P. (1999, Eds.). *Models of Working Memory: Mechanisms of Active Maintenances and executive control*. Cambridge: Cambridge University Press.

Mizumoto, A., & Takeuchi, O. (2008). Exploring the driving forces behind TOEIC scores: Focusing on vocabulary learning strategies, motivation, and study time. *JACET Journal*, 46, 17–32.

Mochizuki, A. (1999). Language Learning Strategies Used By Japanese University Students. *RELC Journal*, 30 (2), 101-113.

Moir, J. & Nation, I. S. P. (2002). Learners' use of strategies for effective vocabulary learning. *Prospect*, 17 (1), 15-35.

Morgan, D. L., (2013). *Integrating qualitative and quantitative methods: A pragmatic approach*. London: Sage publications.

Morgan, J. & Rinvolucri, M. (2004, 2nd Eds.). *Vocabulary*. Oxford: Oxford University Press.

Morse, J. M. & Niehaus, L. (2009). *Mixed Method Design: Principles and Procedures*. London: Routledge.

Muñoz, C. (2012). Explicit Learning in Second Language Acquisition. In C. A. Chapelle (Ed.). *The Encyclopedia of Applied Linguistics*. Boston: Wiley-Blackwell.

Nagy, W., Herman, P., & Anderson, R. (1985). Learning Words from Context. *Reading Research Quarterly*, 20 (2), 233- 253.

Nassaji, H. (2002) Schema theory and knowledge-based processes in second language reading comprehension: A need for alternative perspectives, *Language Learning*, 52, 439-481.

Nassaji, H. (2004). The relationship between depth of vocabulary knowledge and L2 learners' lexical inferencing strategy use and success, *Canadian Modern Language Review*, 61 (1), 107-134.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston, MA: Heinle & Heinle.

Nation, I. S. P. (1993). Using dictionaries to estimate vocabulary size: Essential, but rarely followed procedures. *Language Testing*, 10 (1), 27-40.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63, 59-82.

Nation, I. S. P. (2007). Fundamental issue in modelling and assessing vocabulary knowledge. In H. Daller, J. Milton & J. Treffers-Daller (Eds.). *Modelling and Assessing Vocabulary Knowledge* (35-43). Cambridge: Cambridge University Press.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, 31 (7), 9-13.

Nation, I. S. P. & Wang, K. (1999): Graded readers and vocabulary. *Reading in a Foreign Language*, 12, 355-380.

Newson, R. (2002). Parameters behind “nonparametric” statistics: Kendall’s tau, Somers’ D and median differences. *Stata Journal*, 2 (1), 45–64.

Norris, J. M. & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50, 417-528.

Norusis, M. J. (2012). *IBM® ’ SPSS® ’ Statistics 19 Guide to Data Analysis*. Pearson-Prentice Hall.

Nunan, D. (2005). Classroom Research. In E. Hinkel (Ed.). *Handbook of Research in Second Language Teaching and Learning* (224-240). Mahwah, NJ: Lawrence Erlbaum.

Nyikos, M. & Fan, M. (2007). A review of vocabulary learning strategies: focus on language proficiency and learner voice. In A. D. Cohen & E. Macaro (Eds.). *Language Learner Strategies* (251-273). Oxford: Oxford University Press.

O’Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. London: Sage.

O’Donnell, M. E. (2012). Second Language Learners’ Use of Marginal Glosses. *Foreign Language Annals*, 45 (4), 543–563.

O’Malley, J. M. & Chamot, A. (1990). *Learning Strategies in second language acquisition*. Cambridge: Cambridge University Press.

Orme, J. G. & Combs-Orme, T. (2009). *Multiple regression with discrete dependent variables*. Oxford Univ. Press, USA.

Overstreet, M. (1998). Text enhancement and content familiarity: The focus of learner attention. *Spanish Applied Linguistics*, 2, 229–258.

Oxford, R. (1990). *Language Learning Strategies*. Heinle & Heinle Publishers

Oxford, R. & Nyikos, M. (1989). Variables affecting choice of language learning strategies by university students. *The Modern Language Journal*, 73, 291-300.

Oxford, R. & Shearin, J. (1994). Language Learning Motivation: Expanding the theoretical framework. *Modern Language Journal*, 78, 12-28.

Pacheco, J. (2004). *A Comparison of Textual Input Enhancement and Explicit Rule Presentation in Secondary One English as a Secondary Language Classes*. Master Thesis, Concordia University Montreal, Canada.

Papagno, C. & Vallar, G. (1992). Phonological short-term memory and the language of novel words: the effect of phonological similarity and item length. *Quarterly Journal of Experimental Psychology*, 44 (A), 47-67.

Paribakht, T. S. & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary development. In J. Coady & T. Huckin (Eds.). *Second Language Vocabulary Acquisition* (174-200). Cambridge: Cambridge University Press.

Paribakht, T. S. & Wesche, M. (1999). Reading and "incidental" L2 vocabulary acquisition. *Studies in Second Language Acquisition*, 21 (02), 195–224.

Pellicer-Sánchez, A. (2012a). Automaticity and speed of lexical access: acquisition and assessment. Doctoral dissertation. School of English, University of Nottingham.

Pellicer-Sánchez (2012b). Vocabulary and Reading. In C. A. Chapelle (Ed.). *The Encyclopedia of Applied Linguistics*. Boston: Wiley-Blackwell.

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do *Things Fall Apart*? *Reading in a Foreign Language,* 22 (1), 31-55.

Peters, A. M. (1985). Language segmentation: Operating principles for the perception and analysis of language. In D. I. Slobin (Ed.). *The cross-linguistic study of language acquisition-vol. 2: Theoretical issues* (1029-1067). Hillsdale, NJ: Erlbaum.

Peters, E. (2007b). Manipulating L2 learners' online dictionary use and its effect on L2 word retention. *Language Learning & Technology*, 11 (2), 45–67. Retrieved from http://llt.msu.edu/vol11num2/peters/default.html

Peters, E. (2012). Learning German formulaic sequences: The effect of two attention-drawing techniques. *The Language Learning Journal*, 40, 65–79.

Peters, E., Hulstijn, J. H., Sercu, L., & Lutjeharms, M. (2009). Learning L2 German Vocabulary Through Reading: The Effect of Three Enhancement Techniques Compared. *Language Learning,* 59 (1), 113-151.

Phakiti, A. (2010). Analysing Quantiative Data. In B. Paltridge & A. Phakiti (Eds.). *Research Methods in Applied Linguistics* (39-49). London & New York: Continuum.

Pienemann, M. (2003). Language Processing Capacity. In C. J. Doughty & M. H. Long (Eds.). *The Handbook of Second Language Acquisition* (679-714). Blackwell Publishing.

Pope, C.; Ziebland, S. & Mays, N. (2000). Analysing qualitative data. *BMJ*, 320 (7227), 114-116.

Pulido, D (2003). Modelling the role of second language proficiency and topic familiarity in second language incidental vocabulary acquisition through reading. *Language Learning*, 53 (2), 233-284.

Pulido, D. (2007). The Effects of topic familiarity and passage sight on L2 Lexical Inferencing and Retention through Reading. *Applied Linguistics,* 28 (1), 66-86.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, 52, 513-536.

Ramberg-Beyer, Lis (2013). *Timeless Themes*. Lindthardt og Ringhof.

Rashtchi, M. & Aghili, H. (2014). Computerized Input Enhancement versus Computer-assisted Glosses: Do They Affect Vocabulary Recall and Retention? *Theory and Practice in Language Studies*, 4 (8), 1665–1674.

Rasinger, S. M. (2008). *Quantitative research in linguistics: An introduction*. London: Continuum.

Rasinger, S. M. (2010). Quantitative methods: Concepts, frameworks and issues. In L. Litosseliti (Ed.). *Research Methods in Linguistics* (). London: Continuum.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, 10, 355-371.

Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.). *Vocabulary in a second language: Selection, acquisition, and testing* (209-227). Amsterdam: John Benjamins.

Read, J. & Chapelle, C.A. (2001). A framework for second language vocabulary assessment. *Language Testing*, 18 (1), 1–32.

Reber, A. (2003). Implicit learning. In L. Nadel (Ed.). *Encyclopedia of Cognitive Science*, Vol. 2. London: Nature Publishing, 486-91.

Rebuschat, S., Hamrick, P. Riestenberg, K., Sachs, R., Ziegler, N. (2015). Triangulating Measures of Awareness: A Contribution to the Debate on Learning without Awareness. *Studies in Second Language Acquisition*, 37, 299 – 334.

Révész, A. (2009). Task complexity, focus on form, and second language development. *Studies in Second Language Acquisition*, 31 (3), 437-470.

Révész, A. & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, 35 (1), 31-65.

Révész, A. (2014). Towards a fuller assessment of cognitive models of task-based learning: Investigating task-generated cognitive demands and processes. *Applied Linguistics*, 35 (1), 85-92.

Reynolds, B. L. (2012). Comments on Anthony Bruton, Miguel García López, and Raquel Esquiliche Mesa's "Incidental L2 Vocabulary Learning: An Impracticable Term?" *TESOL Quarterly*, 46 (4), 812-816.

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly*, 10 (1), 77-89.

Richards, J. C. & Schmidt, R. W. (2002). *Longman Dictionary of Language Teaching and Applied Linguistics*. Longman Publishing Group.

Richards, K. (2003). *Qualitative inquiry in TESOL*. Houndmills, Basingstoke: Palgrave Macmillan.

Richards, K. (2009). Interviews. In J. Heigman & R. A. Croker (Eds.). *Qualitative Research in Applied Linguistics* (182-199). Basingstoke: Palgrave Macmillan.

Richards, L. (2009, 2nd edition). *Handling Qualitative Data*. London: Sage.

Richmond, K. (2009). *Inside Reading. The Academic Word List in Context*. Oxford: Oxford University Press.

Rieder, A. (2003). Implicit and explicit learning in incidental vocabulary acquisition. *Views*, 12 (2), 24-39.

Riley, G. (1993). A story structure approach to narrative text comprehension. *The Modern Language Journal*, 77 (4), 417-432.

Roberts, C. W. (Ed.) (1997). *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum Associates.

Robinson, P. (1995). Attention, memory, and the “noticing” hypothesis. *Language Learning*, 45, 283-331.

Robinson, P. (2001). *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, 22 (1), 27-57.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching*, 41 (1), 1-32.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics*, 45, 193–214.

Roby, W. B. (1999). What's in a gloss? *Language Learning and Technology*, 2 (2), 94-101.

Roediger, H. & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1 (3), 181–210.

Roediger, H. & Marsh, E. (2005). The positive and negative consequences of multiple-choice testing. *Journal of experimental psychology. Learning, memory, and cognition*, 31 (5), 1155-1159.

Rott, S. (1999). The effect of exposure frequency on intermediate learners' incidental voabulary acquisition and retention through reading. *Studies in Second Language Acquisition*, 21, 589-619.

Rott, S. (2005). Processing glosses: a qualitative exploration of how form-meaning connections are established. *Reading in a Foreign Language*, 17 (2), 95-124.

Rott, S. (2007). The effect of frequency of input-enhancements on word learning and text comprehension. *Language Learning*, 57 (2), 165–199.

Rubinstein, J. S., Meyer, D. E. & Evans, J. E. (2001). Executive control of cognitive processes in task switching. *Journal of Experimental Psychology: Human Perception and Performance*, 27 (4), 763. American Psychological Association.

Rumsey, D. J. (2011). *Statistics for dummies*. John Wiley & Sons.

Russell, V. (2012). Learning complex grammar in the virtual classroom: A comparison of processing instruction, structured input, computerized visual input enhancement, and traditional instruction. *Foreign Language Annals*, 45 (1), 42-71.

Saldaña, J. (2013). *The Coding Manual for qualitative researchers*. London: Sage.

Salkind, N. J. (2008, 3rd ed.). *Statistics for People Who Think They Hate Statistics*. London: Sage.

Salkind, N. J. (2012). *Statistics For People Who Think They Hate Statistics: Excel 2010 Edition*. London: Sage.

Saragi, T., Nation, I. S. P., & Meister, F. (1978). Vocabulary learning and reading. *System*, 6 (2), 72-78.

Scarcella, R., & Zimmerman, C. (1998). Academic Words And Gender. *Studies in Second Language Acquisition*, 20 (01), 27-49.

Schmidt, R (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11 (2), 129-158.

Schmidt, R. (1994). Implicit learning and the cognitive unconscious: of artificial grammars and SLA. In N. Ellis (Ed.). *Implicit and Explicit Learning of Languages* (165-209), London: Academic Press.

Schmidt, R. (1995). Consciousness and foreign language learning: a tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and Awareness in Foreign Language Learning*, Honolulu: University of Hawai'I at Manoa, Second Language Teaching and Curriculum Center.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (3-32). Cambridge: Cambridge University Press.

Schmidt, R. & Frota, S. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. Day (Ed.). *Talking to learn: Conversation in second language acquisition* (237-326). Rowley, MA: Newbury House.

Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt and M. McCarthy (Eds.). *Vocabulary: Description, Acquisition, and Pedagogy* (199-227). Cambridge: Cambridge University Press.

Schmitt, N. (1998). Tracking the incidental acquisition of second language vocabulary: a longitudinal study. *Language Learning*, 48 (2), 281–317.

Schmitt, N. (2000). Vocabulary in language teaching. Cambridge: Cambridge University Press.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, 12 (3), 329-363.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. London: Palgrave Macmillan.

Schmitt, N. (2014). Size and Depth of Vocabulary Knoweldge: What the Research Shows. *Language Learning*, 64 (4), 313-951.

Schmitt, N. & Carter, R. (2000). The lexical advantages of narrow reading for second language learners. *TESOL Journal*, 9 (1), 4-9.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18 (1), 55-88.

Schmitt, N., Jiang, X. & Grabe, W. (2011). The Percentage of Words Known in a Text and Reading Comprehension. *The Modern Language Journal*, 95 (1), 26-43.

Scott Long, J. (1997). Regression models for categorical and limited dependent variables. *Advanced quantitative techniques in the social sciences*, London: Sage Publications.

Segalowitz, N., & Hulstijn, J. (2005). Automaticity in bilingualism and second language learning. In J. F. Kroll & A. M. B. De Groot (Eds.). *Handbook of bilingualism: Psycholinguistic approaches* (371-388). Oxford: Oxford University Press.

Service, E. & Craik, F. (1993). Differences between young and older adults in learning a foreign vocabulary. *Journal of Memory and Language*, 32, 608-623.

Shafizadeh, H. & Sajedi, R. (2013). Genre-based Teaching and Genre-Familiarity on EFL Learner's Reading Comprehension. *Journal of Social Issues and Humanities*, 1 (3), 67-70.

Sharwood Smith, M. (1991). Speaking to many minds: On the relevance of different types of language information for the L2 learner. *Second Language Research*, 7 (2), 118-132.

Sharwood Smith, M. (1993). Input enhancement in instructed SLA. *Studies in Second Language Acquisition,* 15 (1), 165-179.

Shin, D., & Nation, P. (2008). Beyond single words: the most frequent collocations in spoken English. *ELT Journal,* 62 (4), 339-348.

Shook, D. J. (1994). FL/L2 reading, grammatical information, and the input-to-intake phenomenon. *Applied Language Learning*, 5 (2), 57-93.

Shook, D. J. (1999). What Foreign Language Recalls about the Input-to-Intake Phenomenon. *Applied Language Learning*, 10 (1/2), 39-76.

Singleton, D (1999). *Exploring the Second Language Mental Lexicon*. Cambridge: Cambridge University Press.

Skehan, P. (1989). *Individual Differences in Second-Language Learning*. London: Edward Arnold.

Skehan, P. (1996). A framework of task-based instruction. *Applied Linguistics*, 17, 38-62.

Skehan, P. & Foster, P. (2001). Cognition and tasks. In P. Robinson (2001). *Cognition and Second Language Instruction* (185-205). Cambridge: Cambridge University Press.

Smith, M. S. (1991). Speaking to many minds: On the relevance of different types of language information for the L2 learner. *Second Language Research*, 7 (2), 118–132.

Sökmen, A. J. (1997). Current trends in teaching second language vocabulary. In N. Schmitt & M. McCarthy (Eds.). *Vocabulary: Description, Acquisition, and Pedagogy* (237-257). Cambridge: Cambridge University Press.

Sonbul, S. (2012). Approaches to Second Language Vocabulary Teaching. In C. A. Chapelle (Ed.). *The Encyclopedia of Applied Linguistics*. Boston: Wiley-Blackwell.

Song, M. (2007). Getting learners' attention: Typographical input enhancement, output, and their combined effects. *English Teaching*, 62, 193-215.

Stevick, E. (1976). *Memory, meaning and method*. Rowley, MA: Newbury House.

Stewart, J. & White, D. (2011). Estimating Guessing Effects on the Vocabulary Levels Test for Differing Degrees of Word Knowledge. *TESOL Quarterly*, 45 (2), 370-380.

Stewart, J., Batty, A. & Bovee, N. (2012). Comparing Multidimensional and Continuum Models of Vocabulary Acquisition: An Empirical Examination of the Vocabulary Knowledge Scale. *TESOL Quarterly*, 46 (4). 695-721.

Stewart, J., & White, D. A. (2011). Estimating guessing effects on the Vocabulary Levels Test for different degrees of word knowledge. *TESOL Quarterly*, 45 (2), 370-380.

Stewart, R. (2014). Do Multiple-Choice Options Inflate Estimates of Vocabulary Size on the VST? *Language Assessment Quarterly*, 11 (3), 271-282.

Stewart, R. & Cross, T. (1993). A field test of five forms of marginal gloss study guide: An ecological study. *Reading Psychology: An International Quarterly*, 14 (2), 113-139.

Strauss, A. & Corbin, J. (1998). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. London: Sage.

Strike, K. A. (2006). The Ethics of Educational Research. In J. L. Green, G. Camilli & P. B. Elmore (Eds.). *Handbook of Complementary Methods in Education Research* (57-73). Mahwah, NJ: Lawrence Erlbaum

Swain, M. (2005). The output hypothesis: theory and research. In E. Hinkel (Ed.). *Handbook of Research in Second Language Teaching and Learning* (471-483). Mahwah, NJ: Lawrence Erlbaum Associates.

Swan, M. (1997). The influence of the mother tongue on second language vocabulary acquisition and use. In N. Schmitt & M. McCarthy (Eds.). *Vocabulary: Description, Acquisition, and Pedagogy* (156-180). Cambridge: Cambridge University Press.

Swanborn, M. S. L., & De Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research*, 69, 261-285.

Svensson, H. L. (2012). *Pure guesswork? Lexical inferencing as a strategy for comprehending and learning new foreign language vocabulary while reading*. Doctoral Dissertation. Department of English, Germanic and Romance Studies, University of

Copenhagen.

Tabatabei, O. & Shams, N. (2011). The effect of multimedia glosses on online computerized L2 text comprehension and vocabulary learning of Iranian EFL learners. *Journal of Language Teaching and Research*, 2 (3), 714-725.

Tashakkori, A. & Creswell, J. (2007). The new era of mixed methods. *Journal of Mixed Methods Research*, 1 (1), 3-7.

Tashakkori, A. & Teddlie, C. (2010, 2nd ed.). *Sage Handbook of Mixed Methods in Social and Behavioral Research*. London: Sage.

Tavakoli, H. (2013). A Dictionary of Language Acquisition: A Comprehensive Overview of Key Terms in First and Second Language Acquisition. Rahnama Press.

Taylor, A. (2002). *A meta-analysis on the effects of L1 glosses on L2 reading comprehension*. Unpublished doctoral dissertation, Purdue University. Cited in Taylor, A. (2006). The effects of CALL versus traditional L1 glosses on L2 reading comprehension. CALICO Journal, 23 (2), 309-318.

Taylor, A. (2006). The effects of CALL versus traditional L1 glosses on L2 reading comprehension. *CALICO Journal*, 23 (2), 309-318.

Taylor, A. (2010). Glossing is sometimes a distraction: Comments on Cheng and Good (2009). *Reading in a Foreign Language*, 22 (2), 353-354.

Tekmen, E. & Daloğlu, A. (2006). An investigation of incidental vocabulary acquisition in relation to learner proficiency level and word frequency. *Foreign Language Annals*, 39 (2), 220-243.

Thornbury, S. (2002). *How to teach vocabulary*. Essex: Pearson Education Limited.

Tomlin, R., & Villa, V. (1994). Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition*, 16, 183–203.

Truscott, J. (1998). Noticing in second language acquisition: A critical review. *Second Language Research*, 14 (2), 103-135.

Tseng, W., Dörnyei, Z. & Schmitt, N. (2006). A new approach to assessing strategic learning: The case of self-regulation in vocabulary acquisition. *Applied Linguistics*, 27 (1), 78-102.

Tseng, W., & Schmitt, N. (2008). Toward a model of motivated vocabulary learning: A structural equation modeling approach. *Language Learning*, 58 (2), 357-400.

Ushida, E. (2005). The Role of Students' Attitudes and Motivation in Second Language Learning in Online Language Courses. *CALICO Journal*, 23 (1), 49-78.

VanPatten, B. (1990). Attending to form and content in the input. *Studies In Second Language Acquisition*. 12 (3), 287-301.

VanPatten, B. (1996). *Input Processing and Grammar Instruction: Theory and Research*. Norwood, NJ.: Ablex.

VanPatten, B. (2000b). Processing instruction as form-meaning connections: Issues in theory and research. In J. F. Lee & A. Valdman (Eds.). *Form and meaning in language teaching* (43–68). Boston: Heinle & Heinle.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, 22 (02), 217–234.

Vidal, K. (2011) A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, 61 (1), 219-258.

Waring, R. (1997). Graded and extensive reading: Questions and answers. *The Language Teacher*, 21, 7–13.

Waring, R. (2002). Scales of Vocabulary Knowledge in Second Language Vocabulary Assessment. Retrieved from http://www.robwaring.org/papers/various/scales.htm

Waring, R., & Nation, P. (2004). Second language reading and incidental vocabulary learning. *Angles On The English-Speaking World*, 4, 11-23.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, 15, 130-163.

Watanabe, Y. (1997). Input, intake, and retention. *Studies in Second Language Acquisition*, 19 (3), 287–307.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, 27, 33-52.

Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language*, 20 (2), 232-245.

Webb, S. (2012). Depth of Vocabulary Knowledge. In C. A. Chapelle (Ed.). *The Encyclopedia of Applied Linguistics*. Boston: Wiley-Blackwell.

Webb, S. & Nation, I. S. P. (2012). Computer-Assisted Vocabulary Load Analysis. In C. A. Chapelle (Ed.). *The Encyclopedia of Applied Linguistics*. Boston: Wiley-Blackwell.

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review-Revue Canadienne Des Langues Vivantes*, 53 (1), 13-40.

White, J. (1998). Getting the learners' attention: A typographical input enhancement study. In C. Doughty & J. Williams (Eds.). *Focus on form in classroom second language acquisition* (85-113). Cambridge: Cambridge University Press.

Williams, J.N. (1999). Memory, attention, and inductive learning. *Studies in Second Language Acquisition*, 21 (1), 1-48.

Williams, J. N. (2004). Implicit learning of form-meaning connections. In B. VanPatten, J. Williams, S. Rott, & M. Overstreet (Eds.). F*orm-meaning connections in second language acquisition* (203-218). Mahwah, NJ: Lawrence Erlbaum Associates.

Williams, J. N. (2005). Learning without awareness. *Studies in Second Language Acquisition*, 27 (2), 269-304.

Williams, R. (2006). Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *Stata Journal*, 6 (1), 58–82.

Williams, J. & Evans, J. (1998). What kind of focus and on which forms? In C. Doughty & J. Williams (Eds.). *Focus on form in classroom second language acquisition* (139-155). Cambridge: Cambridge University Press.

Winke, P. M. (2013). The effects of input enhancement on grammar learning and comprehension. *Studies in Second Language Acquisition*, 35 (02), 323–352.

Winter, G. (2000). A Comparative Discussion of the Notion of 'Validity' in Qualitative and Quantitative Research. *The Qualitative Report*, 4 (3/4). Retrieved from http://www.nova.edu/ssss/QR/QR4-3/winter.html

Wong, W. (2001). Textual Enhancement and Simplified Input: Effects on L2 Comprehension and Acquisition of Non-meaningful Grammatical Form. *Applied Language Learning*, 13 (2), 17-45.

Wong, W. (2003). Textual Enhancement and Simplified Input: Effects 0n L2 Comprehension and Acquisition of Non-meaningful Grammatical Form. *Applied Language Learning*, 3 (2), 7-45.

Xu, X. (2010). The effects of glosses on incidental vocabulary acquisition in reading. *Journal Of Language Teaching And Research*, 1 (2), 117-120.

Yoshida, M. (2008). Think aloud protocols and type of reading task: The issue of reactivity in L2 reading research. *Selected Proceedings of the 2007 Second language Research Forum* (199-209). Somerville, MA: Cascadilla Proceedings Project.

Yoshii, N, (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning and Technology*, 10 (3), 85-101.

Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *Canadian Modern Language Review-Revue Canadienne Des Langues Vivantes,* 57 (4), 541-572.

Zandieh, Z. & Jafarigoharm M. (2012). The Effects of Hypertext Gloss on Comprehension and Vocabulary Retention under Incidental and Intentional Learning Conditions*. English Language Teaching*, 5 (6), 60-71.

Zhang, X. (2013). The I don't know option in the Vocabulary Size Test. *TESOL Quarterly*, 47 (4), 790-811.

Zimmerman, C. (1997). Historical trends in second language vocabulary instruction. In J. Coady & T. Huckin (Eds.). Second language vocabulary acquisition (5-20). Cambridge: Cambridge University Press.

ONLINE SOURCES

www.academics.wellesley.edu [On the Man-Whitney U test] http://www.wellesley.edu/Psychology/Psych205/anova.html; accessed on 1/2, 2012

British Educational Research Association - Revised Ethical Guidelines for Educational Research. (2004) = http://www.bera.ac.uk/files/guidelines/ethical1.pdf, accessed on 15/6, 2011

ESRC National Centre for Research Methods = http://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/modules/mod5/9/, accessed on 22/7, 2013

www.laerd.com [On the Kendall's tau correlation coefficient] https://statistics.laerd.com/premium/sroc/spearmans-rank-order-correlation-in-spss-8.php, accessed on 14/9, 2015

www.laerd.com [On ordinal logistic regression in SPSS] https://statistics.laerd.com/premium/olr/ordinal-logistic-regression-in-spss-21.php, accessed on 13/9, 2013

Laufer, B. & Levitzky-Avidad CATSS test online = http://catss.ga/

Meara, P. (2005). _Lex vocabulary tests v2.0. http://www.lognostics.co.uk/vlibrary/index.htm Princeton University Library. Data and Statistical Services. [On regression calculations] http://dss.princeton.edu/online_help/analysis/interpreting_regression.htm, accessed on 20/8, 2013

UCLA: Statistical Consulting Group. [On logistic regression coefficients] = http://www.ats.ucla.edu/stat/stata/dae/logit.htm, accessed on 19/6, 2013.

UCLA: Statistical Consulting Group. [On pseudo R-squareds] http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm, accessed on 9/9 2013.

UCLA: Statistical Consulting Group. [On tobit analysis] http://www.ats.ucla.edu/stat/stata/dae/tobit.htm, accessed 8/9, 2015

www.spss.com [On interpreting correlation coefficient results] http://spssx-discussion.1045642.n5.nabble.com/correlation-and-regression-for-ordinal-and-nominal-dependent-td5471812.html, accessed on 4/6, 2013

Statistics.Stackexchange.com [On the phi-coefficient when interpreting correlation coefficient results] http://stats.stackexchange.com/questions/23938/how-to-correlate-ordinal-and-nominal-variables-in-spss, accessed on 7/9, 2015

www.sas.com [On tied values] http://v8doc.sas.com/sashtml/stat/chap47/sect13.htm, accessed on 2/5, 2013

www.talkstats.com [On effect sizes] http://www.talkstats.com/showthread.php/2063-S.O.S.-effect-size-indexes-for-nonparametric-tests, accessed on 20/6, 2013